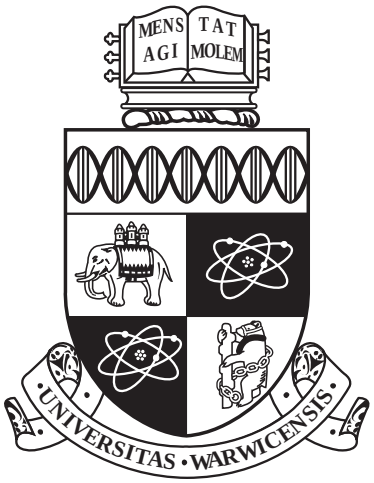


Modelling Techniques for the study of Molecular Self-Organisation

by

Sara Fortuna

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

THE UNIVERSITY OF WARWICK

Department of Chemistry

July 2010

What could we do with layered structures with just the right layers? What would the properties of materials be if we could really arrange the atoms the way we want them? They would be very interesting to investigate theoretically. I can't see exactly what would happen, but I can hardly doubt that when we have some control of the arrangement of things on a small scale we will get an enormously greater range of possible properties that substances can have, and of different things that we can do.

(Richard P. Feynman - annual meeting of the American Physical Society at the California Institute of Technology - December 29th 1959)

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Acknowledgments

Amicitiae nostrae memoriam spero sempiternam fore. (Cicero)

In primis I gratefully acknowledge Dr. Alessandro Troisi, group leader and research supervisor of my postgraduate studies for his guidance, advice and feedback, and the collaborators of this thesis work: Catheline A. L. Colard and Dr. Ir. Stefan A. F. Bon, who performed the experiments reported in Chap. 3, both for the useful discussions on the studied system and for the fruitful collaboration; Dr. David L. Cheung, for suggestions, feedback and comments, in particular regarding Chap. 4. In addition, we thank Prof. Michael P. Allen for suggestions and inputs regarding the work of both Chap. 5 and Chap. 4.

We gratefully aknowledge the members of the Troisi Group directly involved in the project this thesis work is part of, namely "Modelling Techniques for Molecular Self-Assembly" funded by The Leverhulme Trust. We are grateful to Dr. Natalia Martsinovich for useful discussions on the theory of self-assembly of two-dimensional systems, explored in Chap. 4, with particular focus on the TPA system, studied in Chap. 6; Dr. Arijit Bhattacharyay, developer of the data augmentation algorithm (outlined in Appendix A.1), for fruitful discussions in the early stages of the development of the agent based program presented in Chap. 5; and Konrad Diwold, for useful discussion on object oriented programming.

We thanks the other members of the Troisi group: Daniel R. Jones for comments and discussions throughout the whole doctoral activity, and David P. McMahon for carefully proof-reading this thesis. In addition we are thankful to Dr. David Quigley, Dr. Rebecca Notmann, Dr. Christian Diedrich, for inputs and feedback, again throughout the whole doctoral activity.

We finally thank the computing staff who supported this research: Jason Noone, Jaroslaw Zachwieja (aka grok), and Dr. Matthew Ismail. The Computing facilities have been provided by the Centre for Scientific Computing, University of Warwick.

Declaration and Inclusion of Material from a Prior Thesis

Part of the material contained in this thesis have been adapted from the following publications, based on the author's own thesis work:

Chapter 3

1. S. Fortuna, C. A. L. Colard, A. Troisi and S. A. F. Bon - *Packing patterns of silica nanoparticles on surfaces of armored polystyrene latex particles*, Langmuir, 2009, 25 (21), pp 12399-12403.

Chapter 4

2. S. Fortuna, D. L. Cheung and A. Troisi - *Hexagonal lattice model of the patterns formed by hydrogen-bonded molecules on the surface*, J. Phys. Chem. B, 2010, 114 (5), pp 18491858

Chapter 5

3. S. Fortuna and A. Troisi - *An artificial intelligence approach for modeling molecular self-assembly: Agent Based simulations of rigid molecules*, J. Phys. Chem. B, 2009, 113 (29), pp 9877-9885.
4. S. Fortuna and A. Troisi - *Agent Based algorithm for the study of molecular self-organisation*, Proceedings of the Sixth European Conference on Complex Systems, 2009, p24.

Chapter 6

5. S. Fortuna and A. Troisi - *Agent based modelling for the 2D molecular self-organization of realistic molecules*, J. Phys. Chem. B, in press.
6. S. Fortuna and A. Troisi - *Agent Based modelling of realistic molecules* - Proceedings of the ECCS'10 European Conference on Complex Systems, 2010, in press.

In addition, Chapter 1 and Chapter 2 include material from all the cited sources, and the experimental work of Chapter 5 has been done by C. A. L. Colard, under the supervision of S. A. F. Bon.

The author also confirm that this thesis has not been submitted for a degree at another university.

Abstract

In this thesis we develop computational techniques for modelling molecular self-organisation. After a short review of the current nanotechnological applications of molecular self-assembly and the main problems encountered in modelling the self-organised behaviour of chemical systems, we introduce a set of methods, from both chemistry and complexity science, for the prediction of self-assembled structures, with particular focus on Monte Carlo (MC) based methods.

We apply the MC method to two systems of experimental interest. First we model the silica nanoparticles on the surface of spherical polystyrene latex droplets, synthesised by the S. Bon Group at the University of Warwick, as a set of soft spheres on a spherical surface, to study their packing patterns as a function of the broadening of the nanoparticle size distribution. Then we develop a hexagonal lattice model for the study of the two-dimensional self-organisation of planar molecules capable of complementary interactions, to study their phase diagrams as a function of the strength of their complementary interactions and bonding motif. In both cases, the phases are characterised using a number of order parameters. We show that these simplified models are able to reproduce the experimental observations.

We then develop an Agent Based (AB) algorithm, traditionally used for the study of complex systems, for the modelling of molecular self-organisation. In this algorithm, an agent is identified with a stable portion of the system under investigation. The agents can then evolve following a set of rules which include elements of adaptation (new configurations induce new types of moves) and learning (past successful choices are repeated), in order to drive the system towards its lowest energy configuration. We first apply the method to the study of the packing of a set of idealised shapes, then we extend it to the study of a realistic system. The latter is achieved by linking the AB algorithm to an available molecular mechanics code, in order to calculate the interaction energies of atomistic models. In both cases we compare the AB result with that of MC based methods, showing that for all the systems studied, the AB method consistently finds significantly lower energy minima than the MC algorithms in less computing time. Finally, we show how the AB algorithm can be used as a part of the protocol to calculate the phase diagram of a rigid organic molecule (1,4-benzene-dicarboxylic acid or TPA) with less computational effort than standard techniques.

Abbreviations

2D	2-Dimensional
3D	3-Dimensional
AB	Agent Based
AI	Artificial Intelligence
BH	Basin-Hopping
CA	Cellular Automata
CG	Coarse-Grained
DA	Data Augmentation
DFT	Density Functional Theory
GA	Genetic Algorithm
H-bond	Hydrogen bond
IC	Integrated Circuit
LJ	Lennard-Jones
MC	Monte Carlo
MD	Molecular Dynamics
MM	Molecular Mechanics
PBC	Periodic Boundary Conditions
TPA	terephthalic acid

1 INTRODUCTION

We will have to see that
we are the natural expressions of a deeper order.

(Stuart Kauffman, At Home in the Universe, 1995)

SELF-ORGANISATION is the process by which a system increases its order without an external driving force. The process of self-organisation is ubiquitous in nature: ants self-organise into colonies, people into societies, cells into living beings. In chemistry, self-organisation generally refers to the formation of ordered molecular aggregates.

1.1 Self-Organisation in Chemistry

Self-organised structures in chemistry and biochemistry include cell membranes [1], virus capsids [2], and liquid crystals [3]. The units that form the self-assembled structures, or building blocks, can have any kind of shape (for example see Fig. 1.1(a)) and their size can range from that of a nanometer sized single molecule to nanoscale or microscale colloidal particles [4]. The driving forces that promote the self-assembly are the same non-covalent interactions of supramolecular chemistry [5, 6] such as Van der Waals, hydrogen bonds (H-bonds), electrostatics, and hydrophobic interactions.

Self-organised structures are thermodynamically stable, as they tend to adopt a configuration that corresponds to their thermodynamic minimum [8, 7]. In addition, they are often kinetically labile, as they can rapidly explore the available configurational space. These two characteristics make the self-organising process a phenomenon that takes advantage of the reversibility of the non-bonded interactions which, under the right experimental conditions, drives the system towards a unique product [9].

Wales [7, 10] proposed that the energy landcape associated with many self-organisation processes is that of a palm-tree, as shown in Fig. 1.1(b), typical of

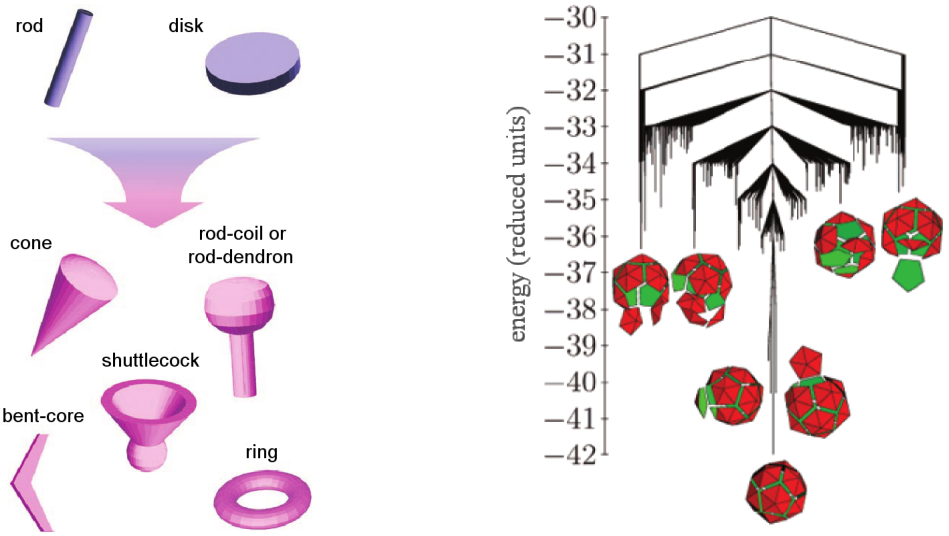


Figure 1.1: (a) Possible shapes for liquid crystal building blocks, adapted from Ref.[3]. (b) Disconnectivity graph for an icosahedral shell composed of twelve pentagonal pyramids, figure adapted from Ref. [7]. Low-lying minima are illustrated near to their corresponding branches in the graph. The energy is in reduced units. This is a method to visualise in two dimensions the multidimensional free energy surface associated with the process. This particular structure for the disconnectivity graph identifies an efficient structure seeker, but in a typical simulation it is easy to get trapped in a relative minimum.

an efficient structure seeker, in which the global minimum of the potential energy surface (represented by the longer vertical line, in the graph of Fig. 1.1(b)) is kinetically accessible as there are no high energy barriers connecting the other local minima (represented by the other vertical lines). Wales also showed how a change in the building blocks can change the picture [10], generating other minima and/or making the global minimum less accessible.

Experimentally, it has been shown that changing the building blocks can even move the global minimum towards a different structure. In fact, as the shape or the strength of the interactions between building blocks changes, different assemblies can emerge, which are often difficult to determine *a priori*. For example, the structures shown in Fig. 1.2 are produced by the assembly of analogous molecules, each containing a rigid core with two hydrophilic groups at each end of the core, plus a hydrophobic side chain. In this example, the structure obtained by the self-assembly of a number of identical molecules, depends on the size of the side

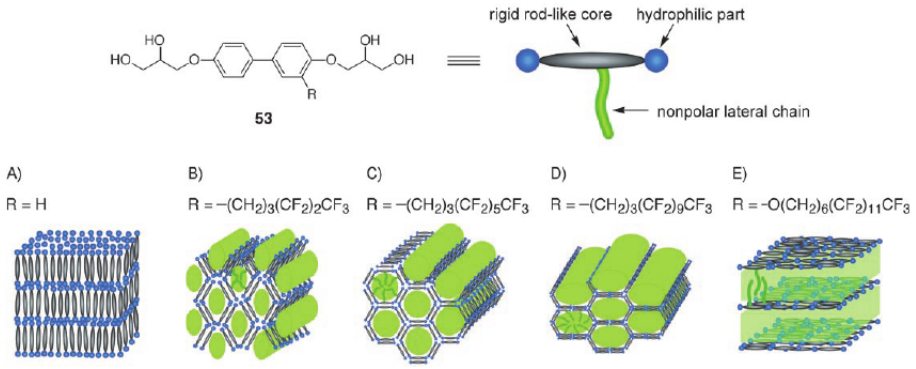


Figure 1.2: Liquid-crystalline phases formed from amphiphilic molecules containing a rigid rod-like core and nonpolar lateral chains; A) smectic A, B) rectangular columnar, C) hexagonal columnar, D) laminated nematic, E) biaxial smectic A. Figure taken from Ref. [3].

chain [3]. The building blocks can also be modified by environmental changes, which leads to different structures. This characteristic is exploited in the study of dynamic materials [11] which can change their structure as a response to external stimuli (e.g. peptide based materials [12] and functionalised nanoparticles [13]).

In integrated circuit (IC) design [14], the self-assembly process constitutes an alternative to current lithographic techniques [15]. In fact, in IC fabrication there are benefits, such as higher speed and less energy consumed per computing function, in decreasing the size of the electronic components [15]. It seems that the lithography based IC fabrication is approaching its limit and, even if the current chips printed by this technique are cheap, in the nanometer-scale self-assembly seems a cheaper alternative. Molecular electronics is being explored for high density nanoelectronic devices [14, 16] where the open problem is how to wire the components together, or better, how to design components capable of self-assembly in the desired fashion. Current research is focused on nanotubes and nanowires [17], and on two-dimensional (2D) self-assembled ordered structures [18]. In particular, the self-assembly of planar molecules on metallic surfaces is of great interest [19, 20, 21, 22] as several systems are capable of forming regular patterns, which in turn can be functionalised through the adsorption of other molecules [23, 24, 25].

In a new field termed “nanomedicine”, assembled architectures and particulate systems are used for diagnostic and targeted drug delivery [26]. Compounds like polymeric micelles [27, 28], dendrimers [29], polymeric nanoparticles [30], protein cages [31] are currently being studied to reach this target. Self-assembled cages

are of particular interest considering that a large number of new drug candidates are not water soluble [26] and the carriers can be engineered such that they can assemble and disassemble as a function of the environmental conditions. In addition, functionalised nanoparticles that can assemble or disassemble, or change their shapes in response to external stimuli can be used for biosensing [32], assay labelling [13], and bioimaging [33, 34].

Deeper knowledge of molecular self-assembly may even impact fields such as medicine, where understanding the process of the formation of amyloid fibres [35], associated with Alzheimer and spongiform encephalopathies, or virus capsids [2] may be a route towards new cures by inhibiting their assembly [36, 37].

In this context, the design of the necessary building blocks and the identification of the correct experimental conditions to obtain a structure with the desired properties are among the most challenging problems. In fact, even though we may have detailed knowledge of the interactions between the building blocks involved, considering a set of molecules and predicting the lowest energy configuration they can form is still challenging [38, 39]. Several methods have been developed for the study of molecular self-organisation, and in the following chapter we will overview them with particular focus on the Monte Carlo (MC) method, the basis of the majority of the work presented in this thesis. In the last two chapters we will propose an alternative approach for the study of molecular self-organising systems by treating them as complex systems.

1.2 Molecular Systems as Multiscale Complex Systems

There is no unique definition of complex systems [40], as this is an umbrella term used to describe systems in which a huge number of relationships can generate new patterns and whose collective behaviour is not easily derivable from the sum of its components behaviours. Self-organised systems, such as colonies, societies, living beings, and molecular aggregates are examples of complex systems and share a set of characteristics. All these systems are formed by identical, or similar, components. The components are related by a number of interactions, which can be either linear or nonlinear, simple or complicated. The sum of these interactions leads to patterns, or *emergent phenomena* (i.e., the self-organisation itself). Given the number of relationships involved, the behaviour of complex systems is difficult to predict *a priori*. For example, looking at a cell we cannot predict *a priori* that an ensemble of cells will produce a living being. In a similar fashion looking at a molecule it is not possible to determine *a priori* which structure a set of molecules will form. The observed patterns are in fact a consequence of the *cooperative effects*

acting among the components, which can contribute to enhance or elide a certain behaviour.

Characteristics that define a complex system, other than cooperative effects, are *hierarchical organisation* and *multiscale properties*. Chemical systems possess these characteristics, as the process of self-organisation takes place at several length-scales and the molecules often come together in supramolecular structures that in turn self-assemble into larger structures. An example is the self-assembly of polypeptides known to form fibres [41, 42]. They interact in the direction of the fibre elongation through several H-bonds and on the other directions thanks to Van der Waals interactions. The polypeptides first form unidimensional fibrils and only at longer timescales will the fibrils self-assemble into fibres. The driving force of the fibre formation is the sum of the Van der Waals interactions acting between the sides of the molecules which is the result of the cooperative effects between the polypeptides. We therefore observe the formation of hierarchical structures at different timescales and different length-scales. A hierarchical spatial organisation is by definition a multiscale process.

We therefore propose in this work that self-organising chemical systems can be studied with the tools of complexity science.

1.3 Aim of this Work

The aim of this thesis is to bind together concepts and techniques from complexity science with the traditional computational chemistry tools in order to study molecular self-organisation.

The main aspect imported from complexity science regards the identification of the simplest features a system needs to possess in order to be able to express a determined behaviour. In this context, we will assume that self-organisation is an emergent phenomenon due to the interactions among a collection of particles and find the minimum number of relations able to give a certain pattern. In self-assembly this can be achieved either (i) with a minimal model to represent the system under investigation and/or (ii) with a minimal representation of its dynamics. In this thesis we will investigate both of these aspects.

First, we will explore one example in which the available simplified models provide an adequate representation of the experimental system (synthesised in Stefan Bon's group). We will model polymeric droplets coated with silica nanoparticles as a set of spheres on a spherical surface, verifying that simplified models sometimes suffice to describe experimental systems and are useful for collecting additional information on scales not reachable by current experimental techniques.

Then we will develop a hexagonal lattice model, for the study of the 2D self-assembly of planar molecules on surfaces. We will show that this symplified model is able to reproduce many of the patterns encountered in experimental systems.

Finally, we will move onto the key aspect of this thesis work, by introducing a new modelling paradigm in the field of molecular simulations. This will be done through the implementation of a rule-based algorithm, namely an Agent Based (AB) algorithm, first proposed as a lattice algorithm [43]. First the AB algorithm will be developed for the study of off-lattice systems, then, in order to be able to recover the system-specific knowledge, we will link the AB algorithm to an available code to calculate properties of a realistic system.

1.4 Thesis Outline

This thesis is organised as follows: in Chap. 2 we will overview models and methods that can be used for the study of molecular self-organisation, with particular focus on the MC method. The latter will be used in Chap. 3 to study the packing patterns of spherical silica nanoparticles on the surface of spherical polystyrene latex droplets, and in Chap. 4 to study the 2D patterns formed by planar molecules capable of complementary interactions on metallic surfaces, with an hexagonal lattice model. In Chap. 5 we will introduce our AB model for the study of molecular self-organisation of a set of idealised shapes, and in Chap. 6 we will extend the AB model for the study of realistic system. Chap. 7 concludes the thesis.

2 MODELLING TECHNIQUES AND THEORETICAL BACKGROUND

.. it is requisite that its parts mutually depend upon each other
both as to their form and their combination,
and so produce a whole by their own causality ..

(Immanuel Kant, The Critique of Judgement, 1892)

We introduce here the background knowledge for the modelling of self-assembling chemical systems. We first discuss the model systems used to describe the building blocks of the molecular self-organisation, then, after some basic definitions of statistical physics, we will briefly review simulation methods used both in chemistry and in complexity science.

2.1 Model Systems

The building blocks of molecular self-assembly usually consist of molecules containing between 10 and 1000 atoms. They can be defined at many levels of physical accuracy: as occupied lattice sites, as idealised shapes, or as point masses (atoms) held together by suitable interatomic potentials. Simple lattice models have been very useful, before the advent of the computers, to calculate analytically the first phase diagrams [44]. The advent of computers however, together with the increase of available computational power, allowed for the study of more complicated systems: from simple systems composed of spherical particles, up to detailed atomistic models. In this thesis, we will treat the building blocks as rigid bodies, without internal degrees of freedom, focusing our attention on the intermolecular forces driving the self-assembly process.

2.1.1 Lattice Models

The simplest model we can think of, for the description of a set of identical particles, is that of the lattice gas model [46], first introduced in 1952 by Lee and Yang to

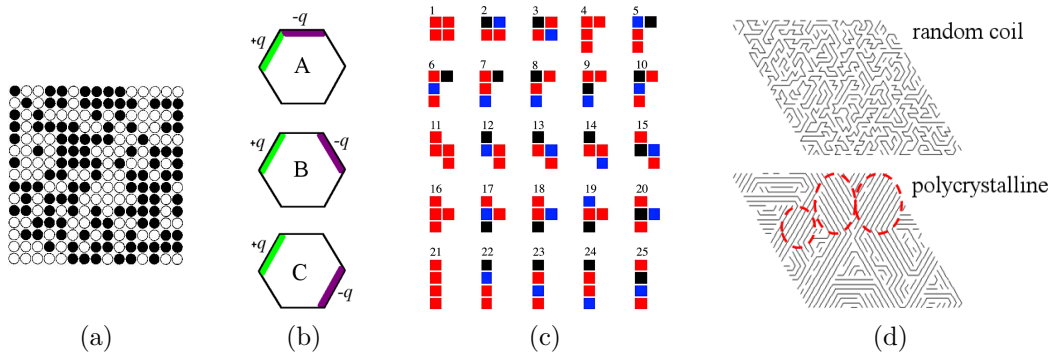


Figure 2.1: (a) Gas-lattice model: occupied sites are black and empty sites are white. (b) Hexagonal particles with two charged edges, figure adapted from Ref.[45]. (c) Tetromino particles occupy multiple lattice sites, in this case each site can be neutral (red), positively (blue) or negatively (black) charged, figure adapted from Ref.[43]. (d) Lattice polymers in random coil (top) and polycrystalline (bottom) conformations.

study phase transitions analytically [47]. In a lattice gas model, particles are represented as occupied sites on a lattice, often a square lattice as in Fig. 2.1(a). The system is described by a Hamiltonian that takes into account the Van der Waals interactions between particles:

$$H = -\epsilon \sum_{\langle i,j \rangle} n_i n_j \tag{2.1}$$

where ϵ is the interaction strength, $n_i = 0$ if the cell i is empty, $n_i = 1$ if it is occupied, and the summation is taken over adjacent cells. The lattice gas model, derived from the Ising model [46] used for the study of ferromagnetism and in which each lattice site can assume the states $s_i = \pm\frac{1}{2}$, can be converted into the Ising model by the linear transformation $n_i = \frac{s_i+1}{2}$. Therefore the model is isomorphic to the Ising model, and all the results for the Ising model also apply to the lattice gas model. The Ising model, proposed and studied analytically in one dimension by Ising in 1925 [48] and extended to 2 dimensions by Onsanger in 1944 [49], is the first of a small set of analytically solvable models for the study of phase transitions [44].

Among the extensions of the simple lattice gas model, there are Potts models [50] in which the lattice sites can assume a number of states larger than two, and models in which the particles have an internal structure. Examples of the latter include Ice-Type models on the square lattice [44], the Poker Chip model on

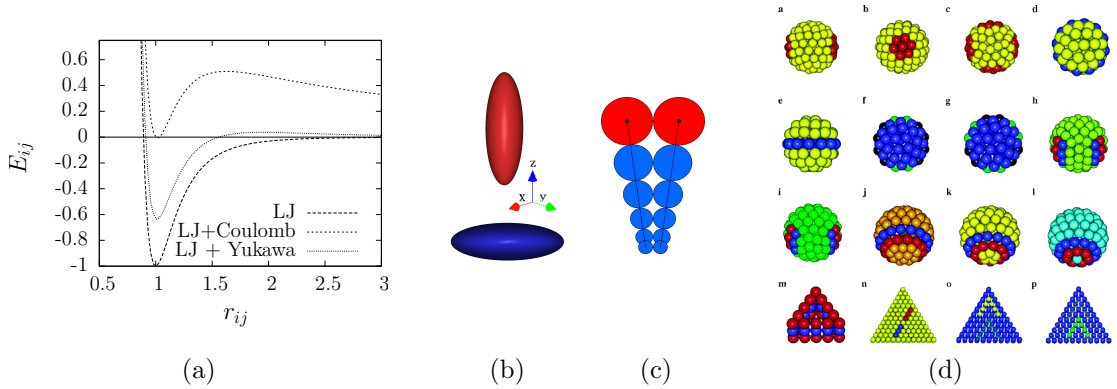


Figure 2.2: (a) Spherically symmetric potentials. (b) Gay-Berne particles: elongated (top) and discotic (bottom). (c) Two amphiphilic particles with hydrophilic headgroups (red) and hydrophobic tails (blue) to model micelle formation, taken from Ref. [61]. (d) Model with distinguishable particles, taken from Ref. [62].

the hexagonal lattice [51], and the model we will introduce in Chap. 4 where the hexagonal building blocks (Fig. 2.1(b)) interact with a Hamiltonian able to describe both Van der Waals interactions and directional H-bonds. It is also possible to include long range interactions, for instance a dipolar interaction $H_{\mathrm{D}} = D \sum_{ij} \frac{n_i n_j}{r_{ij}^3}$ as in Ref. [52]. Models of this type are rarely analytically solvable.

Other extensions allow particles to occupy multiple lattice sites, as in the case of tetromino particles [43, 53]. Tetromino particles are particles that occupy four lattice sites and are the first particles whose self-assembly has been studied with an AB model (see Ref. [43] and Sec. 2.4.3). In Ref. [43], each site composing the tetromino particles could be neutral, positively or negatively charged as shown in Fig. 2.1(c). Other examples include lattice polymer models (Fig. 2.1(d)). In these models, monomers occupy a single lattice site and they can be characterised as hydrophobic or polar, like in the HP model proposed in 1989 by Lau and Dill for the description of proteins [54], or can interact through more complicated terms to model, for instance, the Van der Waals-like interaction or the H-bond formation between polymers [55]. Lattice models have a long history in the study of polymers [56] and have also been used to study the behaviour of proteins, biomacromolecules [57, 58], and liquid crystals [59, 60].

2.1.2 Rigid-Soft Models

In the continuum space, the sphere is the simplest particle model we can think of. Simple impenetrable spheres that cannot overlap are described by a *hard sphere potential* [63]:

$$E_{ij}^{\text{hard}} = \begin{cases} 0 & \text{if} \quad |r_{ij}| \geq r_i + r_j \\ \infty & \text{if} \quad |r_{ij}| < r_i + r_j \end{cases} \tag{2.2}$$

where r_{ij} is the distance between the centers of the spheres with radii r_i and r_j respectively. This was one of the first systems studied using computer simulations by the pioneers of the field [64, 65, 66, 67].

Another system of interest, first approached in the paper of Rosenbluth and Rosenbluth [65], is a system composed of soft spheres described by a Lennard-Jones (LJ) like potential (Fig. 2.2(a)) of the form:

$$E_{ij}^{\text{LJ}\alpha-2\alpha} = \epsilon_0 \left[\left(\frac{r_{\text{eq}}}{r_{ij}} \right)^{2\alpha} - 2 \cdot \left(\frac{r_{\text{eq}}}{r_{ij}} \right)^{\alpha} \right] \tag{2.3}$$

where ϵ_0 is the depth of the potential well, r_{eq} is the equilibrium distance between them, and α is a parameter describing the "softness" of the potential. Large values of α give a harder potential, in which particles are less likely to compenetrate, and *viceversa*. For example, the argon atoms described by a LJ potential with $\alpha = 6$ have very soft boundaries and this is consistent with the observation that the boundaries are defined by their electron clouds and allow for compenetration, on the other hand the silica nanoparticles of Chap. 3, being composed of several atoms, are better described with a harder potential (for example $\alpha = 12$).

Charges can be included by addition of a Coulomb term:

$$E_{ij}^{\text{C}} = \frac{1}{4\pi\epsilon} \frac{q_i q_j}{r_{ij}} \tag{2.4}$$

where q_i and q_j are the charges of each particle and ϵ is the electrical permittivity. Alternatively, as in the case of colloidal particles, it is possible to describe the electrostatic interactions through a Yukawa potential (or screened Coulomb potential) [68, 69]:

$$E_{ij}^{\text{Yukawa}} = \frac{A \cdot \exp(-r_{ij}/\xi)}{(r_{ij}/\xi)} \tag{2.5}$$

where ξ is the screening length parameter, and $A = \frac{q_i q_j}{4\pi\epsilon}$. The screening parameter takes into account the fact that colloidal particles often have surface charges and are surrounded by electrolytes which screen the effective Coulomb interactions between colloids.

Hydrophobic terms are more difficult to include, as these interactions are due to the rearrangement of water molecules and H-bonds as two hydrophobic molecules come together. However, they can be modelled for example with a potential of the form $E_{ij}^{\mathrm{hp}} = -\gamma^{-r_{ij}/\lambda}$, where γ is a surface tension constant, and λ the range of the hydrophobic interaction [61].

If the building blocks are not spherically symmetric and anisotropic, as with liquid crystals, it is possible to use the Gay-Berne potential [70]:

$$E_{ij}^{GB} = 4\epsilon \left[\left\{ \frac{\sigma_{\mathrm{xy}}}{r_{ij} - \sigma + \sigma_{\mathrm{xy}}} \right\}^{12} - \left\{ \frac{\sigma_{\mathrm{xy}}}{r_{ij} - \sigma + \sigma_{\mathrm{xy}}} \right\}^{6} \right] \tag{2.6}$$

where both ϵ and σ depends first of all on the orientation of the particles, and their separation vector. In addition, Gay-Berne particles interact with different strengths in the x, y plane with respect to the z direction, and this is taken into account in the functional form of ϵ, and the aspect ratio between the two can be adjusted to give either elongated (Fig. 2.2(d), top) or discotic (Fig. 2.2(d), bottom) particles. The latter aspect is taken into account by the form of σ which depends on the length and breadth σ_{xy} of the particles.

Alternatively, it is possible to describe a non spherical building block using several LJ spheres. For example Tsonchev *et al.* [61] studied the self-assembly of amphiphilic molecules for the study of micelle formation, modelling them as rigid cones composed of spheres which interact with a hydrophobic term plus a Coulombic term (Fig. 2.2(c)). Sciortino [71] studied gel formation of systems composed of patchy particles, and LJ spheres with attractive spots on their surface. Zhang and Glotzer [62] studied several systems of particles composed of different LJ spheres each characterised by a different value of ϵ_0, as shown in Fig. 2.2(d), to understand the effect of the shape and of the different interaction strengths on the self-assembled structures. In Chap. 6 we will test our algorithm on a set of three-dimensional (3D) tetromino-like shapes composed of different arrangements of LJ spheres which can be neutral, positively, or negatively charged.

These simplified models are able to give a general understanding of the behaviour of classes of systems that share the same characteristics. However, in order to get insight into the self-assembly of realistic molecules, atomistic models and coarse-grained (CG) models should be used.

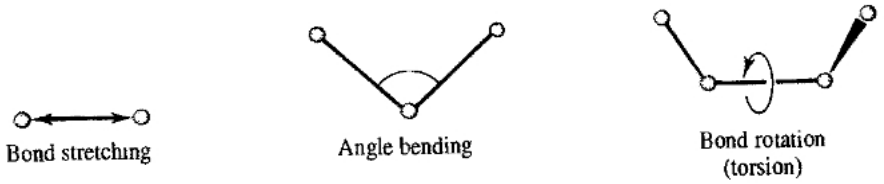


Figure 2.3: Force-field contributions. Taken from Ref. [72].

2.1.3 Molecular Mechanics Based Models

Molecular mechanics (MM) models rely on a classical treatment of particle-particle interactions. Molecules in MM are represented as particles connected by a set of "springs", which represent the chemical bonds. Each particle can represent either a single atom, as in atomistic models, or a group of atoms, as in CG models. The interaction between the particles are described by a *force-field* [72], which typically includes intramolecular terms such as bending, stretching, and torsions (see Fig. 2.3), and intermolecular terms, such as Van der Waals, electrostatics, and H-bonding.

The bonding energy is well described by a Morse potential $E_{\text{Morse}}(l) = D_e[1 - e^{-a(l-l_0)}]^2$ with D_e depth of the energy minimum, $a = \omega\sqrt{\mu/2D_e}$ where ω is the frequency of the bond vibration, μ is the reduced mass, l_0 the average bond length, and l the current length. However, as it requires three parameters, usually the harmonic potential is used instead, which requires only two parameters:

$$E_{\text{stretching}}(l) = \frac{k_l}{2}(l - l_0)^2 \tag{2.7}$$

This correspond to the first nonzero term of a Taylor expansion around the equilibrium configuration. Additional terms of the expansion, such as cubic or higher terms, can be included to better reproduce the Morse curve. The bond-angle energy can be computed as:

$$E_{\text{bending}}(\theta) = \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{2.8}$$

where k_θ is the harmonic force constant, θ_0 is the equilibrium angle between 3 particles, θ the current angle. As for the stretching energy, higher order terms can

also be included. The torsional energy can be expressed as:

$$E_{\text{torsion}}(\omega) = \sum_{n=1}^{N} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) \tag{2.9}$$

where V_n gives an indication of the height of the energy barriers between the n minima, γ is the phase factor which determines the position of the energy minima, and the number of terms N can be adjusted to take into account the number of expected minima. In addition other terms can be included, such as out-of-plane bendings and cross-terms to couple two degrees of freedom [72].

The basic functional form of the intermolecular potential is usually based either on a LJ potential (described in the previous section) or on a Buckingham potential

$$E_{\text{buckingham}} = Ae^{-Br} - Cr^{-6} \tag{2.10}$$

which has a softer repulsive component with respect to the LJ potential. Electrostatic contributions can be taken into account as charge-charge interactions, through a Coulomb term as in Eq. 2.4, or as bond dipoles. In addition, other terms can be included such as H-bonds energies, or quadrupoles moments.

All the functions used to describe both intra- and intermolecular energies, depend on a set of parameters that can be either obtained from quantum mechanical calculations or from experimental data. Functions together with their parameter set completely define a force-field. In the case of atomistic models, well established force-fields include MM2/MM3 [73, 74] first proposed by Allinger in 1977 for small organic molecules, and Amber [75, 76] proposed by by Weiner and co-workers in 1984 [75] for proteins and nucleic acids.

If a full atomistic representation is not feasible, it is possible to *coarse-grain* the system [77, 78], considering groups of atoms as a single unit. This allows the study of larger systems at longer timescales. The parameterisation of CG models is usually performed *via* the comparison with atomistic models. Examples of CG potentials include the MARTINI potential [79], for the study of amphiphilic molecules where, on average, four heavy atoms are represented by a single CG unit. Other CG models are often used to describe proteins [80], in which each amino acid can be represented as a particle (or bead), or polymers [81], in which every repetitive unit is represented as a bead. In both cases, side chains can be taken into consideration as additional beads.

The fact that CG models represent groups of atoms make them closer to the experimental systems than the hard/soft potentials introduced in the previous sec-

tion, as they can be derived from and mapped back to an atomistic representation [81]. The possibility of a mapping between different levels of approximation is a key feature of the multiscale models that will be introduced in Sec. 2.3.3.

2.2 Some Statistical Mechanics Definitions

A system at equilibrium can be in one of many possible microstates s. A property Q will assume a value $Q(s)$, depending on the current microstate of the system. However, experimentally we measure Q in a finite time interval τ performing a time average:

$$Q_{\text{exp}} = \frac{1}{\tau}\int_0^\tau Q[s(t)]dt \tag{2.11}$$

as the microstate of a system $s(t)$ can change over the observation time τ. In simulations time is discrete, and for molecular dynamics (MD) simulations (see Sec. 2.3.1) we can measure the same property as:

$$Q_{\text{MD}} = \langle Q\rangle_t = \frac{1}{n\Delta t}\sum_{i=1}^{n} Q[s(i\Delta t)] \tag{2.12}$$

As this is not always possible nor necessary, we make use of the *ergodic theorem* that states that, at the thermodynamical equilibrium, averages taken over time on a given system are the same as averages taken over many replicas of the same system with the same macroscopic characteristics. The replicas of the system under investigation with similar macroscopic parameters form a *statistical ensemble*. A number of ensembles are known from statistical thermodynamics [82, 83] and are commonly used: (i) the microcanonical ensemble or NVE ensemble, in which all the systems have the same volume V, the same number of particles N, and are thermally isolated with the same total energy E; (ii) the canonical ensemble or NVT ensemble, in which all the systems are in thermal equilibrium with a large heat reservoir or heat bath large enough to keep the temperature T constant, and (iii) the grand canonical ensemble, which is in thermal contact with a bath and allows for exchange of particles.

The ergodic theorem is of great practical use, as it allows us to run simulations on a number M of replicas of a system, and average over observations taken at the same simulation time t':

$$\langle Q\rangle_M = \frac{1}{M}\sum_{m=0}^{M} Q[s_m(t')] \tag{2.13}$$

In addition, given a statistical ensemble, if we define p_s as the probability of a replica being in a certain state s, we can average over states writing:

$$Q = \langle Q \rangle_s = \sum_s Q_s p_s \tag{2.14}$$

Eq. 2.14 is the form used in MC simulations (Sec.2.3.2).

2.3 Simulation Methods in Chemistry

Once the building block representation has been chosen, it is necessary to choose a theoretical method to describe the system properties. Several methods have been developed to study molecular systems and, traditionally, each simulation method is bound to a unique time/length-scale, as shown in Fig. 2.4. At the extreme of small and large scales there are methods that do not rely on the previously introduced models: quantum chemical calculations and continuum models. At small time/length-scales quantum chemical studies have been successfully used to identify the lowest energy dimers, trimers, and tetramers thought to be the supramolecular building blocks of larger structures [84, 85, 86]. At the other extreme, where very large time/length-scales prevail with respect to the atomic constituents, as in fluid dynamics [87] or the simulation of mechanical properties, continuous models can be employed. Between the two scales there are the two methods used for the study of self-organisation of ensembles of molecules: MD and MC.

In this thesis, we will mainly focus on MC-based methods. However, for completeness, in this section we will also briefly overview the MD method.

2.3.1 Deterministic Methods: Molecular Dynamics

The MD simulation method [88] is based on Newton's second law or the equation of motion, $\vec{F}_i = m_i \vec{a}_i$, where $\vec{F}_i$ is the force exerted on a particle i with mass m_i, and $\vec{a}_i$ is its acceleration. The force can be expressed as the gradient of the potential energy (E): $\vec{F}_i = -\vec{\nabla}_i E$, where E is represented by one of the force-fields introduced in Sec. 2.1. Integration of the equations of motion yields a trajectory that describes the positions and velocities of the particles as they vary with time. From this trajectory, the average values of properties can be determined through Eq. 2.12. The method is deterministic: once the positions and velocities of each atom are known, the state of the system can be predicted at any time in the future or the past.

MD simulations are extremely useful to follow the dynamic of a system, as

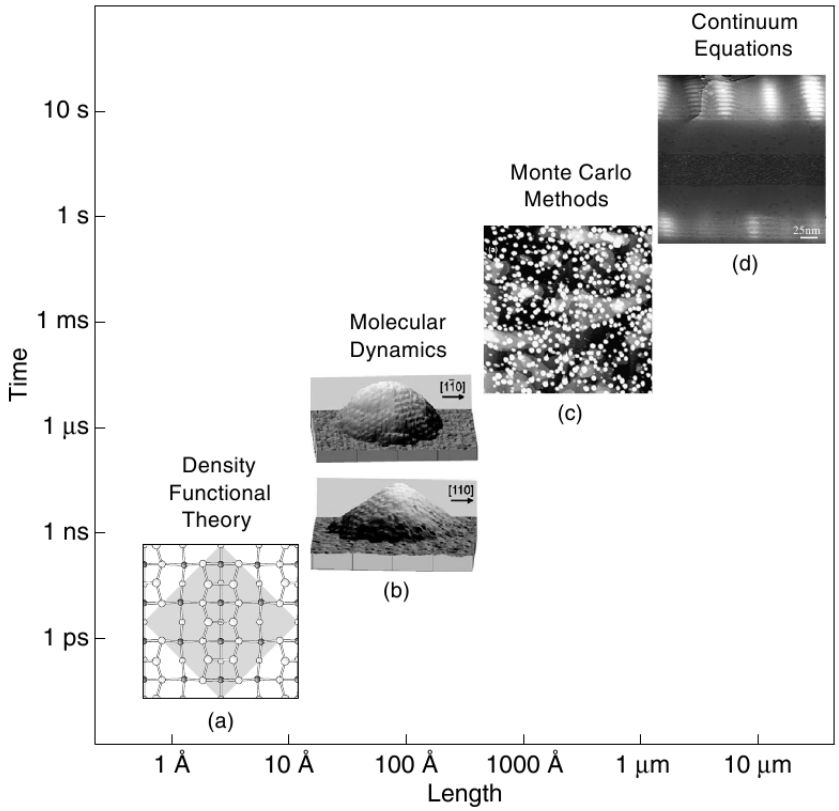


Figure 2.4: Simulation methods in chemistry. On the x-axis the system length-scale, on the y-axis the timescale. From this plot it is possible to see that every technique is bound to a unique time/length-scale. Taken from Ref. [81].

time is explicitly included, allowing for the calculation of time-dependent properties. Usually MD simulations relax fast towards the nearest local minima and are not suitable for overcoming high energy barriers between configurations, because of the very small time/length-scales they can reach [89]. However, if the target configuration of a system is known, they are a useful tool to study its equilibrium behaviour.

Many implementations of the method have been developed in order to escape the local energy minima, and therefore reduce the computational time by artificially speeding up the exploration of the configurational space. Strategies like hyperdynamics [90], temperature accelerated methods [91], metadynamics [92], activation-relaxation techniques [93, 94] might be used to overcome this problem. For example in hyperdynamics [90] a bias potential can be used to raise the energy of the system in regions other than that of the transition state, but this method has never been used for cluster formation or self-assembly. Similarly the tempera-

ture accelerated method, in which part of the simulation is carried out at higher temperature, has not been used for these scopes. Metadynamics [92] is one of the methods of choice to study crystal nucleation [95, 96], as with this method it is possible to fill the explored potential wells with Gaussians [97], in order to flatten the surface and explore the whole configurational space. This method has been successfully used by Quigley *et al.* to study water [95] and calcite nucleation [96]. The activation-relaxation technique [93, 94] allows the system to explore neighbouring local minima by deforming the system (activation) to let it escape a local minimum and then the system is pushed into an adjacent minimum (relaxation). Mousseau explored the early stages of fibre formation with an activation relaxation technique [98] observing, for short sequences, the formation of parallel dimers, trimers and hexamers.

MD methods are therefore implemented for the detailed study of small groups of molecules, making it possible to explore the transition steps in the very first stages of molecular self-organisation, but to study systems composed of hundreds of molecules they are limited by the short timescales and long computational time required.

2.3.2 Stochastic Methods: Monte Carlo

The MC method is a stochastic method, first proposed by Ulam and Metropolis in 1949 [99] to study systems composed of a large but finite number of particles. The idea behind the MC method [64, 83, 88] is that, in order to estimate Eq. 2.14 without *a priori* knowledge of the functional form of p_s, it is possible to consider the possible states of a system at the equilibrium as a *reversible Markov chain*, and estimate Q as:

$$Q_M = \frac{1}{M}\sum_{i}^{M} Q_i \tag{2.15}$$

where the average is taken over a Markov chain of length M.

A Markov chain is a chain of states in which the transition probability $P(s' \rightarrow s)$ from a state s' to a state s depends only on s', allowing one to write the master equation of the system as:

$$\frac{dp_s}{dt} = \sum_{s'}\Big[p_{s'}(t)P(s' \rightarrow s) - p_s(t)P(s \rightarrow s')\Big] \tag{2.16}$$

where p_s and $p_{s'}$ are the probability of the system being in the state s and s',

respectively. At equilibrium, the steady state approximation ($\frac{dp_s}{dt} = 0$) applies:

$$\sum_{s'} P(s \to s')p_s = \sum_{s'} P(s' \to s)p_{s'} \tag{2.17}$$

and, in order to eliminate limit cycles, in which the system explores cyclically a set of states without necessarily exploring the whole state space, the detailed balance should be satisfied:

$$P(s \to s')p_s = P(s' \to s)p_{s'} \tag{2.18}$$

which can be rewritten as:

$$\frac{P(s \to s')}{P(s' \to s)} = \frac{p_{s'}}{p_s} \tag{2.19}$$

In 1902, Gibbs showed that a system in thermal equilibrium with a reservoir, in the NVT ensemble, can be in one of several possible microstates s with probability:

$$p_s = \frac{1}{Z} e^{-E_s/k_\mathrm{B}T} \tag{2.20}$$

where E_s is the energy associated with the microstate s, k_B is the Boltzmann constant, T is the system temperature, and Z is the partition function of the system. The partition function can then be derived as a normalisation constant, given that $\sum_s p_s = 1$ we get:

$$Z = \sum_s e^{-E_s/k_\mathrm{B}T} \tag{2.21}$$

combining Eq. 2.19, 2.20 and 2.21, the transition probability becomes:

$$\frac{P(s \to s')}{P(s' \to s)} = e^{-\frac{E_{s'} - E_s}{k_\mathrm{B}T}} \tag{2.22}$$

Given the transition probabilities, it is then possible to choose any algorithm to build up the chain of states. We can therefore break up the transition probabilities into two terms, separating the probability $g(s \to s')$ of proposing a new configuration s' from the probability of accepting the proposed configuration $A(s \to s')$, and rewrite Eq. 2.22 as follows:

$$\frac{P(s \to s')}{P(s' \to s)} = \frac{g(s \to s')}{g(s' \to s)} \cdot \frac{A(s \to s')}{A(s' \to s)} \tag{2.23}$$

In the Metropolis algorithm, one sets $g(s \to s') = g(s' \to s)$ (i.e., the probability of proposing a move is the same for opposite moves) and the relation between the

acceptance probabilities must be:

$$A(s \rightarrow s') = A(s' \rightarrow s)e^{-\frac{E_{s'}-E_s}{k_{\mathrm{B}}T}} \tag{2.24}$$

with $0 \leq A(s \rightarrow s') \leq 1$ and $0 \leq A(s' \rightarrow s) \leq 1$, as they are probabilities. As we want to maximise the acceptance ratio, it is convenient to give the larger of the two acceptance probabilities the value 1 and adjust the other. The Metropolis acceptance probability is therefore:

$$A(s \rightarrow s') = \min[1, e^{-\frac{E_{s'}-E_s}{k_{\mathrm{B}}T}}] \tag{2.25}$$

first tested in 1953 for the simulation of a set of 224 rigid spheres in two-dimensions [64].

In general, the Metropolis sampling scheme can be used either for off-lattice and lattice models, and it is useful for systems that do not present high energy barriers between states [100, 101]. For example, in this thesis, the method will be applied to simple systems in both Chap. 3 and Chap. 4, to simulate silica nanoparticles on the surface of a polymeric nanodroplet, as a set of LJ spheres on a spherical surface, and H-bonded molecular networks on a metallic surface, with a hexagonal lattice model, respectively. The Metropolis MC scheme works well for simple systems and it has been proven that systems simulated with MC will converge to the Boltzmann distribution, but the number of simulation steps required to reach equilibrium cannot be determined *a priori* [102], and the simulations can take a very long time. However, it is possible to improve MC simulations for the study of self-assembled structures in several ways.

The first simplest modification of the MC method for structure prediction is simulated annealing [83, 103], inspired by the analogy with physical annealing, where a system, initially at a high temperature disordered state, is slowly cooled down becoming more and more ordered. This is achieved in the MC simulation by varying the system temperature T according to a cooling schedule. In the *annealed MC*, at the beginning of the simulation the acceptance probability $A(s \rightarrow s')$ is very high, leading the system to explore many different configurations by escaping from local minima, then, as the simulation evolves the probability $A(s \rightarrow s')$ becomes smaller and high energy configurations are less likely to be accepted, leading the system to its global minimum. This is the traditional MC-based technique used to solve global minimisation problems [83]. In Chap. 6 we will compare this algorithm with the AB algorithm developed in Chap. 5.

Alternative techniques of global optimisation in molecular systems include basin-hopping (BH) [7], replica exchange (also known as parallel tempering [104]), and the introduction of multiple particles moves into the MC algorithm. In BH [7] a new configuration is generated, its energy is minimised, and is then accepted or rejected based on the Metropolis criterion. In this way the real potential energy surface is transformed into a set of plateaus (= basins) where the barriers between local minima are flattened out. In replica exchange [104, 105], two or more parallel simulations are run at different temperatures and, occasionally, configurations are swapped between the simulations with a probability given by $P_{a,b}^{\text{swap}} = \min\left[1, \exp\left[-\frac{E_b - E_a}{k_{\text{B}}(T_a - T_b)}\right]\right]$ where E_a is the energy of the system a at temperature T_a and E_b is the energy of the system b at temperature T_b.

Multiple particle moves can be introduced in the algorithm defining groups of particles both on geometric [106] or energetic [107, 108] considerations. For example, the algorithm developed by Liu [106] is based on geometric considerations as follows: at every simulation step a pivot point is chosen at random, and a random particle is moved by reflection with respect to the pivot point, when this correspond to an overlap with other particles the other particles are moved with the same method. This is repeated until no overlap occurs anymore. On the other hand, the algorithm implemented by Bhattacharyay [107], the data augmentation (DA) algorithm, is based on energetic considerations: two particles are considered forming a cluster if they are bonded, and they are bonded with a certain probability that depends on their interaction energy. At every DA step, a particle is chosen, a cluster is defined, and it is moved. The cluster move is then accepted or rejected with a modified acceptance probability (see Appendix A.1), in order to keep the detailed balance. Among the algorithms capable of group moves, the DA algorithm [107, 108] is the one with which we will compare the algorithm presented in Chap. 5.

The introduction of group moves in a MC simulation scheme is a key aspect to reach faster convergence [107, 108], as this allows for multiscale moves capable of escaping local energy minima. This aspect will be included in the AB algorithm of Chap. 5 and Chap. 6.

2.3.3 Multiscale Methods

According to Peter and Kremer [81], the term multiscale simulation "refers to methods where different simulation hierarchies are combined and linked to obtain an approach that simultaneously addresses phenomena or properties of a given system at several levels of resolution and consequently on several time- and length-

scales."

The coexistence of several simulation hierarchies can be achieved by introducing multiscale moves into a MC algorithm, as shown in the previous section where we introduced MC algorithms capable of binding different length-scales thanks to group moves. Another way to introduce group moves into a MC algorithm, is to mix it with MD in a hybrid scheme as with mixed MC-stochastic dynamics methods [109]. In this algorithm MC steps are alternated with stochastic dynamics steps, in order to perform group moves, thanks to MD, and reach longer timescales than MD alone without losing the microscopic detail. Even though the possibility of multiscale moves is a key element of the algorithms of Chap. 5 and Chap. 6, the implementation of another aspect of the multiscale modelling is a parallel field of open research: the implementation of strategies to combine different levels of resolution of the system under examination.

The possibility of multiple resolutions currently involves a sequence of simulations going from detailed simulations (e.g. MD with atomistic models) towards CG simulations, in order to extend the approachable time/length-scales with MD simulations. At every step of the procedure, the results of the detailed simulation are used to parametrise the next level of description and, eventually, build up a mapping between the two in order to recollect the microscopic details from the mesoscopic simulations. Procedures of this type have been used for the study of polymers [110, 111]. Hybrid simulations are also used, in which atomistic details are considered for the interesting portions of the system and the rest of the system is described with CG models. In this framework, the adaptive resolution scheme (AdResS), developed by Praprotnik [112], is particularly promising: this is a method in which molecules can enter and exit an atomistic region and adapt their description to the region in which they reside (see Fig. 2.5). However, currently these approaches have not been used for the study of self-organising systems, as they are very preliminary works.

In this thesis, we argue that the way to swap easily among scales with little computational effort relies on rule based models, in particular AB models, described in the next section. We believe that the inclusion of features like adaptive resolution schemes, might form part of the future development of the AB algorithm developed in Chap. 5 and Chap. 6.

2.4 Simulation Methods in Complexity Science

It is important to point out that MC and MD schemes are generally used for one of three tasks: (i) generating low energy (at least metastable local minima) configura-

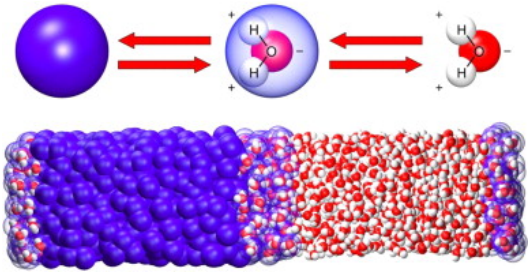


Figure 2.5: Adaptive resolution scheme. Molecules adapt their resolution level "on the fly", water molecules can go from a CG description (left) to an atomistic one (right) and *viceversa*. Figure taken from Ref. [112].

tions; (ii) sampling from the ensemble of configurations available to the system at equilibrium; (iii) finding physical pathways between stable or metastable states, on either the energy or free energy landscape. In this thesis we are mainly concerned with generating low energy configurations and sampling the ensemble configurations of a system at equilibrium, neglecting the last aspect. As a sampling scheme we will use the Metropolis MC scheme throughout all the thesis. However, when generating low energy configurations, regardless if the evolution of coordinates towards those configurations is unphysical, we can take advantage of a number of other methods. Examples includes *artificial intelligence* (AI) based methods, e.g. genetic algorithms (GA), or *rule based models*, such as cellular automata (CA) or AB models.

2.4.1 Genetic Algorithms

A GA is a "search techniques based on the principles of natural evolution" [113, 114, 115] and it is generally used as a global optimiser. The parameters to be optimised are collected into a vector. In a GA every vector is considered an individual, part of a *population*, that represents a trial solution of the problem. The population evolves following bio-inspired rules as follows: at every simulation step (or generation) each individual is ranked according to a *fitness function*, which quantifies the quality of the trial solution, and then selected or discarded following tournament rules in order to create a new population. The new population is created either by a *crossover* operation, in which two individuals selected from the previous generation are mixed to generate *offsprings*, or by *mutation*, in which an offspring is created by a random modification of an individual of the previous generation. The offsprings are then the new starting population for the next simulation step. This model has

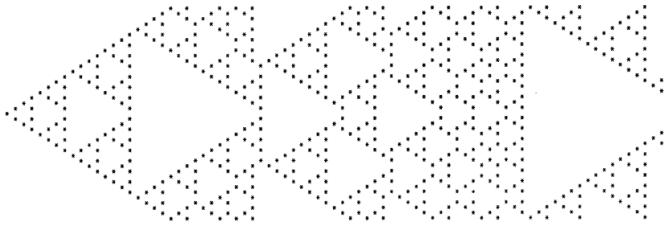


Figure 2.6: One-dimensional CA; the time step increases from left to right and successive time-steps are shown on successive columns; the 1-cells are represented by stars, the 0-cells by blank. Figure adapted from Ref. [120].

been successfully used for cluster geometry optimisation [113], where each cluster corresponds to an individual that, at every simulation step, is randomly rotated and cut into two parts. Two randomly selected parts belonging to two different clusters, are then merged into an offspring, and the offspring are ranked (in this case their fitness is related to their internal energy) and selected with a Metropolis acceptance probability (as Eq. 2.25), in order to pass to the next generation. The method has also been used for crystal structure prediction [116], protein folding [117], and self-assembly [118].

2.4.2 Cellular Automata

CA are lattice rule based models [119], in which each cell can be in one of a finite number of states and can change state following a set of rules based on its own state and the states of the neighboring cells, called the neighbourhood. At each simulation step, the rules are applied to the whole grid and a new generation is created.

For example Fig. 2.6 shows a one-dimensional two-state CA that presents self-similarity and fractal structure [120]. The updating rule is very simple: when the two neighbouring sites have the same value, the site assumes the value 0 independently of its current state, otherwise it assumes the value 1. Analogous patterns can emerge also in 2D CA from a few nonzero initial cells [121]. Wolfram [120] proposed that this and similar models can be interpreted as a crystal growth from an initial seed.

CA have been extensively used to study complex systems from both physical and social sciences. In this context an example is the sandpile model [122], which is defined as follows: given a square lattice, each point of the lattice is associated with an integer number indicating the height of the sandpile. At each time step, the

height of a random point is increased by unity. If the height at one point exceeds a critical value, then a topping event occurs, the height of the point is reduced by 4 units and the height of the 4 neighbouring sites is increased by one. If any of the neighbours is now unstable, the process continues until none of the points of the system exceed the critical value. The size distribution of the avalanches follows a power law distribution, proving that a small perturbation (i.e., adding one single grain to a sandpile) can sometimes lead to very large avalanches. This model has been generalised to represent earthquakes [123] or economic events [124].

CA also have applications in biology, as in the study of population dynamics[125], and in chemistry [126, 127, 128, 129] as it has been extensively proven that the rules are particularly suitable as an alternative to partial differential equation models. CA have been used in chemistry to study reaction kinetics [126], recrystallisation [127], particle nucleation [128] and fluctuations in cluster formation [129]. The capabilities of this modelling paradigm are obviously limited by its discrete nature.

2.4.3 Rule Based Artificial Intelligence: Agents

AB models are rule based models based on the collective behaviour of a set of agents. By definition, "an *agent* is a computer system capable of exchanging information with other agents and its environment, taking decisions and performing autonomous actions" [130]. AB models have traditionally been used in fields like economics [124] or sociology [122, 131] to model the behaviour of complex systems such as stock markets and societies. For example, to model the stock market [124], agents can be identified with traders and, to model social phenomena, like the emergence of hierarchy in the society [131], agents can be identified as warriors. Agents share a set of characteristics: (i) they have a a set of properties, such as a position in space, a pot of money if they are traders, or an amount of "power" if they are warriors [131]; (ii) they evolve in time through a set of actions, for example buying and selling stocks or fighting; (iii) they are goal-oriented, for example traders want to maximise their income and warriors want to maximise their power. In order to reach their goal, agents will decide which action to perform basing their decision on a set of conditions. Conditions and actions form the rules of an AB model. In general, the AB rules are: (i) *nonlinear*, as they can contain if-then conditions or more complicated algorithmic decisions; (ii) *local*, because only the local environment may be considered by the agent before undertaking an action; and (iii) *adaptive*, allowing the rules to evolve in time, to best suit the goal of the simulation. In particular, the adaptation is a characteristic of the AB rules not

present in the CA ones. The adaptability of the rules can have a great effect on the performance of the algorithm, leading it to reach the expected goal with less computational effort, such as less memory storage and/or computing time.

AB models have been introduced in chemistry [132, 133, 134, 43, 135], under the assumption that chemical systems are complex systems and it is therefore possible to describe their complex dynamics with a set of rules. For example, Cartwright *et al.* [132, 133] modelled the enantiomeric crystallisation of molecules letting each agent correspond to a nucleation site of L or D handedness. Each agent is then put into a flux, to study the effect of the stirring on the growing crystals. The model is based on four simple rules: new agents can appear both as new nucleation sites (with random handedness), or derive from an already existing nucleus (therefore with its same handedness), they can then either grow or dissolve. This model shows that, in agreement with the experiments [136], crystals under stirring conditions are all D or L. Bradford and Dill studied the the self-organisation of proteins [134]. In their model, proteins acting as catalysts are identified as agents able to move in space, with moves biased such that each agent is driven towards a region of space rich in its reactant. The end simulation result is that proteins tend to self-assemble into multicomponent aggregates in order to catalyse all the reactions of a chain. Troisi, Wong, and Ratner [43] studied the packing of a set of molecules to find their lowest energy configuration. In their model, an agent is identified with a shape or a group of shapes on a square lattice. At the begginning of the simulation each agent coincides with a rigid shape and, as the simulation proceeds, the agent evolves to represent stable portions of the system. Each agent evolves due to three actions: *move* to a new position of the lattice, *merge* with another agent, and *split* into two agents. In Chap. 5 and 6 we will implement this model in order to describe off-lattice systems of idealised and of atomistic particles.

3 MONTE CARLO SIMULATION OF POLYDISPERSE SPHERES ON A SPHERICAL SURFACE

There is geometry in the humming of the strings,
there is music in the spacing of the spheres.

(Pythagoras, 6th.century BC)

N COLLABORATION with the Bon Group, we studied the packing pattern of silica nanoparticles on the surface of spherical polystyrene latex droplets with MC simulations. The experimental system has been modelled as a set of interacting spheres on a spherical surface. The information supplied by this model has been complementary to the experimental data. We study the effect of the polydispersity of the spherical nanoparticles on the self-assembled structure. We show that broadening of the nanoparticle size distribution has pronounced effects on the self-assembled equilibrium packing structures, with the original 12-point dislocations or grain-boundary scars gradually fading out.

3.1 The Sphere Packing Problem

Packing patterns of identical and non-identical spherical and discotic objects on curved surfaces are often encountered in nature and science. Examples include C60 fullerenes [137, 138], 13-atom cuboctahedral metal clusters [139], S-layer proteins on outer cell membranes [140] which are all formed by the self-assembly of identical building blocks, and the lenses on insect eyes, biomineralized shells on coccolithophorids [141], solid-stabilised emulsion droplets [142] and bubbles [143], made of building blocks of different sizes.

It is well known that the maximum packing density of a single layer of spheres of identical size in an infinite 2D flat plane is achieved when they are arranged into a hexagonal lattice, with each of the spheres having six neighbours. It is also known from the literature [144] that a set of equally sized spheres, or calottes, on a spherical surface cannot form a regular hexagonal packing, due to the positive

curvature of the surface. The determination of the packing geometry of identical spheres or circles onto a spherical surface is often referred to as the Thomson problem [145], generalised later by Tammes [146]. The generic approach is to position the spheres, or circles, as far away from each other as possible, using a repulsive Coulomb potential. In 3D hexagonal HCP or FCC lattices of identical spheres, 12 nearest neighbours of each sphere can be identified. When we look at an isolated cluster of 13 spheres in such a lattice, each of the 12 spheres assembled onto the central one has 5 neighbours (excluding the central sphere). The deviation from having 6 nearest neighbours in a 2D hexagonal packing arrangement on a flat surface is a direct effect of the curvature of the surface. This number is also an exact solution of the Tammes problem. In every spherical system, 12 of these so-called defects must be present, and when the central sphere becomes larger, more neighbours can be accommodated on its surface. Examples include a football or its chemical equivalent, the C60 buckyball [137, 138], which is composed of 12 pentagons and 20 hexagons. Bausch *et al.* [147, 148] investigated very large systems with the assembly of thousands of microspheres of identical size onto emulsion droplets. They showed that the generic rule of 12 defects prevailed in the form of five- and seven-neighbour line defects, or grain-boundary scars.

The influence of size variations on these packing patterns is studied sparsely, and interesting questions arise. First of all it is unknown what happens in systems of intermediate size, and what would be the packing organisation when hundreds of particles are assembled onto a sphere. In addition, as it is yet unknown what characteristics will have a self-assembled equilibrium packing structure if the building blocks are not all the same size. To try to answer these questions, we studied the assembly of silica nanoparticles on the surfaces of small submicrometer-sized droplets of styrene, which for imaging purposes were solidified by free radical polymerisation *via* a so-called Pickering miniemulsion polymerisation process [149, 150]. The packing organisation of such intermediate-sized systems is of practical importance in applications such as the film formation of armoured polymer latexes in waterborne coatings [151] and the fabrication of reinforced permeable supracolloidal structures of submicrometer size with potential use as a nanocuvette for single-molecule spectroscopy measurements [152].

3.2 Synthesis of Coated Nanoparticles

The synthesis and imaging of the nanoparticles have been performed by Catheline Colard, under the supervision of Stefan Bon. They made use of the phenomenon that solid particles can adhere strongly to liquid-liquid or liquid-gas interfaces, as

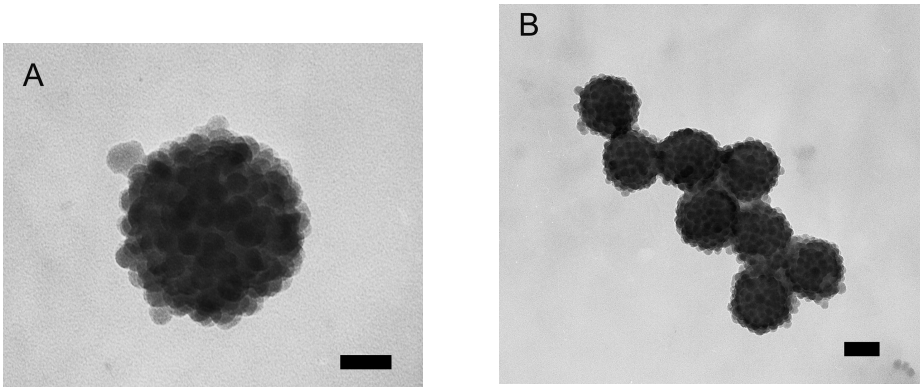


Figure 3.1: TEM images of colloidal silica armoured polystyrene latex particles (batch **1**). Scale bars: (A) 50 nm and (B) 100 nm.

demonstrated over a century ago by Ramsden [153] and Pickering [154]. The driving forces of the process are the minimisation of the interfacial energy [155] and, when charged particles are used, image-charge stabilisation [156]. The Bon experimental group previously developed a Pickering miniemulsion polymerisation protocol [149, 150] in which they decorated submicrometer-sized droplets of monomer (e.g. styrene), with nanosized cylindrical Laponite clay discs, and subsequently polymerise them. They then used the same approach by replacing the Laponite clay discs with Ludox TM-40 silica nanoparticles (waterborne nonviscous sol of 40 wt %). For the silica nanoparticles to wet styrene and allow adhesion to the styrene-water interface, experiments were performed under acidic conditions at a pH of ca. 3.5. This warranted limited surface charges. TEM analysis showed that these nanoparticles have an average particle radius of 12.03 nm with a standard deviation of 1.68 nm.

Armoured polymer latexes were prepared straightforwardly using the Pickering miniemulsion polymerisation procedure. An interesting result was that nanocomposite latexes could be prepared with overall solid contents exceeding 50 wt %. Monomodal particle size distributions of the latexes were obtained, with an increase in polydispersity for concentrated systems. Varying the number of silica nanoparticles produced only limited control of particle size, whereas in Pickering miniemulsion polymerisation using Laponite clay, as a solid stabiliser, control of the particle size was achieved. The latter was possible when the high energy emulsification step (i.e., sonication) generated an equilibrium situation for the partitioning of the clay discs, thereby dictating the total droplets surface area [149].

The underlying explanation for limited control of particle size when silica nanoparticles are used is that the difference in shape (i.e., spheres vs discs) creates a large shift in the concentration of nanoparticles present, which directly affects the rate of nanoparticle adhesion, as can be described by the von Smoluchowski theory [157]. This geometric difference lowers the rate of collision and thus adhesion, making it of the same order of magnitude or slower than the rate of creation of a new liquid-liquid interface in our emulsification step. The diffusion limitation leads to the fusion of partially covered droplets and hence limited control of the particle size upon variation of the relative number of silica nanoparticles.

EM investigation showed that the polystyrene latex spheres, made *via* Pickering miniemulsion polymerisation, were armoured with a layer of silica nanoparticles (Fig. 3.1). The nanoparticles were packed in close proximity on the surface, showing that surface charges were in fact minimal. Because of the constrains of assembly on curved surfaces and a variation in the silica particle size, interesting packing patterns emerged.

To investigate these patterns more in-depth, two batches of armoured polymer latexes using the same recipe were prepared. Their size distribution was 209.70 ± 21.75 nm for batch **1**, and 211.91 ± 31.11 nm for batch **2**. For clarity, we will focus our discussion on the results obtained from batch **1**. Similar results were obtained for batch **2**.

SEM analysis of approximately 25 armoured polymer latex particles showed that the silica particles were arranged in patterns with either 4, 5, 6, or 7 neighbours (Fig. 3.3A), at percentages of 12, 47, 37, or 2.9, respectively (Fig. 3.3C). This clearly deviates from the ideal 12-point dislocations seen in 13-atom 3D hexagonal HCP or FCC clusters of identical spheres, in which the 12 outer atoms each have 5 neighbours, and from larger systems such as C60 showing 12 pentagonal and 20 hexagonal faces. It also deviates from the large assemblies of particles on large droplets investigated by Bausch *et al.* [147, 148], where exclusively 5- and 7-neighbour grain-boundary dislocations were found.

Abkarian *et al.* [158] described the formation of stable buckled faceted polyhedral supracolloidal shapes from bubbles covered with monodisperse polystyrene particles, with particle-to-bubble radius ratios of about 0.1. EM analyses of our silica nanoparticle armoured polystyrene latexes show no clear evidence of such polyhedral shape formation. In Ref. [158] the edges of the polyhedra on armoured bubbles are clearly seen, unlike in our case (Fig. 3.1 and Fig. 3.2A). A key difference is that the volume of the solid- stabilised bubbles decreases after jamming, and this leads to the formation of polyhedra with a larger surface-to-volume ratio.

Nevertheless, one can state that a small (< 200) number of points on a convex surface define a polyhedron, and our observed structures could, in theory, be described and modelled as such. We argue, however, that it is not convenient to map the observed structure into polyhedra because a vast majority of the polyhedra that can be built with a plausible distribution of coordination numbers have shapes that are so distant from that of a sphere that they cannot represent our experimental situation (consider, for example, the elongated nature of the C70 molecule). In other words, the miniemulsion droplets impose a very rigid constraint on the possible arrangement of the silica nanoparticles, and it is better to include this constraint from the very beginning instead of theoretically generating many polyhedra and then removing the less stable ones. Moreover, when polydispersity is taken into account (as we must in our case), there is an enormous number of possible particle size distributions for each particular polyhedron, and one must resort to a statistical exploration of the possible structures (i.e., a MC method) because the enumeration of all cases is not possible (unlike for the monodisperse case).

3.3 Monte Carlo Simulations on the Spherical Surface

To understand and rationalise the newly found packing patterns we performed a series of MC simulations modelling the system as a set of spheres on a spherical surface. In our model, the center of each nanoparticle is located on a spherical surface with radius $R+r_j$, where R is the radius of the droplet and r_j is the radius of nanoparticle j. The nanoparticles are modelled as soft spheres interacting *via* a LJ-like potential of the form

$$E_{ij} = \epsilon_0[(r_{\mathrm{eq}}/r_{ij})^{2\alpha} - 2(r_{\mathrm{eq}}/r_{ij})^{\alpha}] \tag{3.1}$$

where ϵ_0 is the depth of the potential well, $r_{\mathrm{eq}} = r_i + r_j$ is the equilibrium distance between two nanoparticles i and j, and r_{ij} is the actual distance between the two. We used both a standard 6-12 LJ potential ($\alpha = 6$) and a narrower 12-24 LJ potential ($\alpha = 12$). Electrostatic contributions were not taken into account in the main results presented, which is a reasonable approximation for the experimental conditions. All MC simulations were run at $k_{\mathrm{B}}T = 0.01\epsilon_0$, where $k_{\mathrm{B}}T$ is the thermal energy (i.e., in a regime at which the thermal energy of the system is much smaller than the interaction energy between particles).

In the simulation, the particles sitting on the spherical surface are described by the spherical coordinates (θ, ϕ). As this is a non-orthogonal coordinate system, to ensure a uniform sampling of the spherical surface, given a state θ, ϕ, a new state

θ', ϕ' should be proposed with probability:

$$g(\phi \to \phi') = \min\left[1, \frac{\sin\phi'}{\sin\phi}\right] \tag{3.2}$$

and, making use of Eq. 2.23, the transition probability becomes:

$$P\big(\{\theta,\phi\} \to \{\theta',\phi'\}\big) = g(\phi \to \phi') \cdot A\big(\{\theta,\phi'\} \to \{\theta',\phi'\}\big) \tag{3.3}$$

where $A\big(\{\theta,\phi'\} \to \{\theta',\phi'\}\big)$ is the Metropolis acceptance probability (Eq. 2.25). As the two probabilities are independent, the transition probability $P(\phi \to \phi')$ can also be written as a conditional probability:

$$P\big(\{\theta,\phi\} \to \{\theta',\phi'\}\big) = P\big(\{\theta,\phi'\} \to \{\theta',\phi'\} \mid \phi \to \phi'\big) \tag{3.4}$$

and this is the form used in the implementation of this model.

To find the optimum number of nanoparticles assembled in the most favourable packing configuration, a series of simulations were performed by increasing the nanoparticle number of one unit and after each time a new equilibrium configuration was found. To determine the average energy of a system with N assembled nanoparticles, 25 million production MC steps have been performed after an equilibration of the same length. Every simulation has been repeated five times, and the presented results are obtained from their average.

We performed two sets of MC tests in which (i) we validated the theoretical model, by comparing the nearest neighbour distribution between our simulations and the experimental data obtained from the Pickering miniemulsion polymerisations and (ii) we investigated the influence of the polydispersity of the particle size distribution on the self-assembled packing pattern and the influence of the size of the droplet as well as the intercorrelation between these two. Although many studies have been devoted to the packing of polydisperse spheres in 2D and 3D [159, 160] and the packing of spheres on a spherical surface with hard [144], soft [161], and long range repulsive potential [162, 163], the problem of packing polydisperse spheres on a spherical surface has not previously been investigated.

To validate the theoretical model by comparison with experiments, we set the radius of the uncovered droplet equal to 80.79 nm. The radius of the nanoparticles was normally distributed with average radius $r = 12.03$ nm and standard deviation of $\sigma = 1.68$ nm, corresponding to a polydispersity σ/r of 0.14 (as in the experiment). The optimum number of nanoparticles, corresponding to the lowest overall energy, was 219 ± 5 using a standard LJ potential and 206 ± 6 in case of

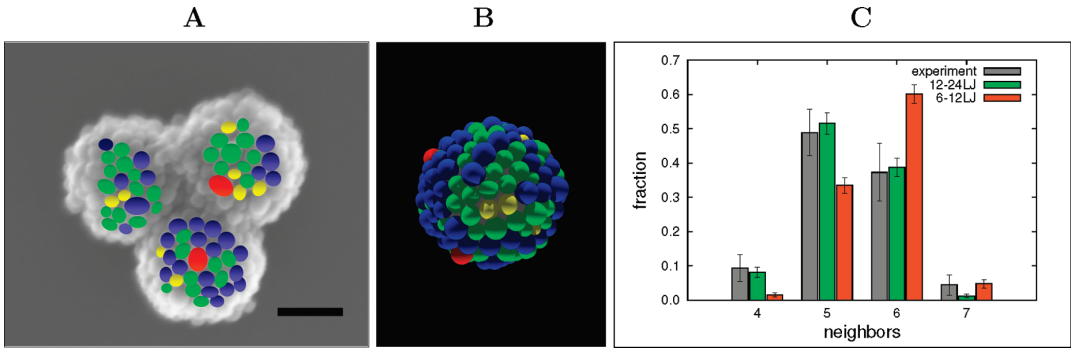


Figure 3.2: Packing patterns of nanosized silica particles on the surface of polystyrene latex particles: (A) Experimental micrograph taken by SEM (scale bar 100 nm). (B) Simulation snapshot of one equilibrium configuration with the 24-12 LJ potential (205 particles). The colours represent the number of neighbours (see text) of each silica nanoparticle: yellow (four), green (five), blue (six), red (seven). (Colours in the micrograph have been added manually.) To compare the experimental result and the simulated data calculated with both the 6-12 LJ and 12- 24 LJ potentials, the fraction of neighbours are compared (C). The histogram shows excellent agreement between the experiment and the simulation with the 12-24 LJ potential. The data has been averaged over two batches. Five simulations have been run for each batch.

a narrowed LJ potential. We classified the particles according to the number of neighbours (both in the simulation and from the experimental micrograph) defining two particles as neighbours if their distance is shorter than the sum of the radii incremented by 15%.

The statistics for the number of neighbours are displayed in Fig. 3.2C, showing excellent correlation between experiment and simulation for the narrower interparticle potential used. This is very reasonable because the 12-24 LJ potential corresponds to a width of the potential energy well of approximately 6% of the nanoparticle radius (i.e., ∼0.7 nm), in agreement with the EM micrograph. The 6-12 LJ potential is too soft (potential energy well width ∼1.7 nm) to account for the packing distances observed in the experiment. We present in Suppl. S.1.1 the results for a number of different intermolecular potentials. It is interesting to notice that the commonly used Yukawa potential [68] (and likely any other long range repulsive potential [162]) for charged colloidal particles would provide a completely different distribution of coordination numbers with a very high fraction of 4- and 5-coordinate nanoparticles. The good agreement between experimental and simulated coordination numbers, once the potential with the correct softness is

determined, is also a good indicator that the observed system has reached thermodynamic equilibrium.

To investigate the influence of the particle size distribution broadening on the self-assembled packing patterns and the influence of the size of the droplet, we simulated a matrix of 5×4 systems using polydispersities σ/r of 0.00, 0.07, 0.14, 0.21, and 0.28 and droplet radii R of 161.58, 80.79, 40.40, and 20.20 nm. Given that the selected average radius r of the nanoparticles is 12.03, the dimensionless system sizes R/r are 13.4, 6.7, 3.4, 1.68 (i.e., from twice to one-quarter of the experimentally observed size). These values allow us to explore the transition from point defects (such as the ones found in systems with $R/r < 1.5$) and the scars (or line defects) described by Bausch *et al.*, who studied the cases of $R/r = 24$ and 44 [148]. To analyse more systematically the nature of defects in the nanoparticle packing, and how it is influenced by the nanoparticle polydispersity, we used the concept of Delaunay tessellation [148, 164]. It consists of a triangulation of the spherical surface of the nanodroplet in which the spherical coordinates of each nanoparticle are identified as a vertex on the surface of a sphere (i.e., we divide the surface into triangular subsections, with each triangle being formed by the nearest three points). With this method, the coordination number of each nanoparticle is defined as the number of triangles that belong to it. In all systems, the sum of the dislocation charges (defined as the difference between the ideal coordination number of 6 and the actual coordination number of the particle) is 12 in order to satisfy the Euler formula. We note incidentally that, unlike the apparently related problem of finding all isomorphs of carbon fullerenes [165], the problem of describing the most stable arrangement of nanoparticles on a nanodroplet cannot be mapped into a relatively simple mathematical problem because the nanoparticles are constrained on a spherical surface and are not free to form a generic convex surface (as carbon fullerenes do).

Fig. 3.3A shows the equilibrium configurations of the system with $R/r = 6.7$ and how the fraction of spheres with coordination number from 4 to 8 changes as a function of polydispersity (Similar results are found for all considered values of R/r). A small polydispersity ($\sigma/r = 0.07$) has no effect on the defect concentration, and the structure, even if more strained, is the same as in the monodisperse case. As the polydispersity increases, the 5-7 disclinations increase in number, and dislocations 4 and 8 appear. Interestingly, the appearance of 4-8 defects is simultaneous with a slight excess of coordination number 4 (as this keep the net dislocation charge equal to 12). Systems with different R/r values behave similarly: in the monodisperse case, we observe 12 point defects or scars, as seen in the

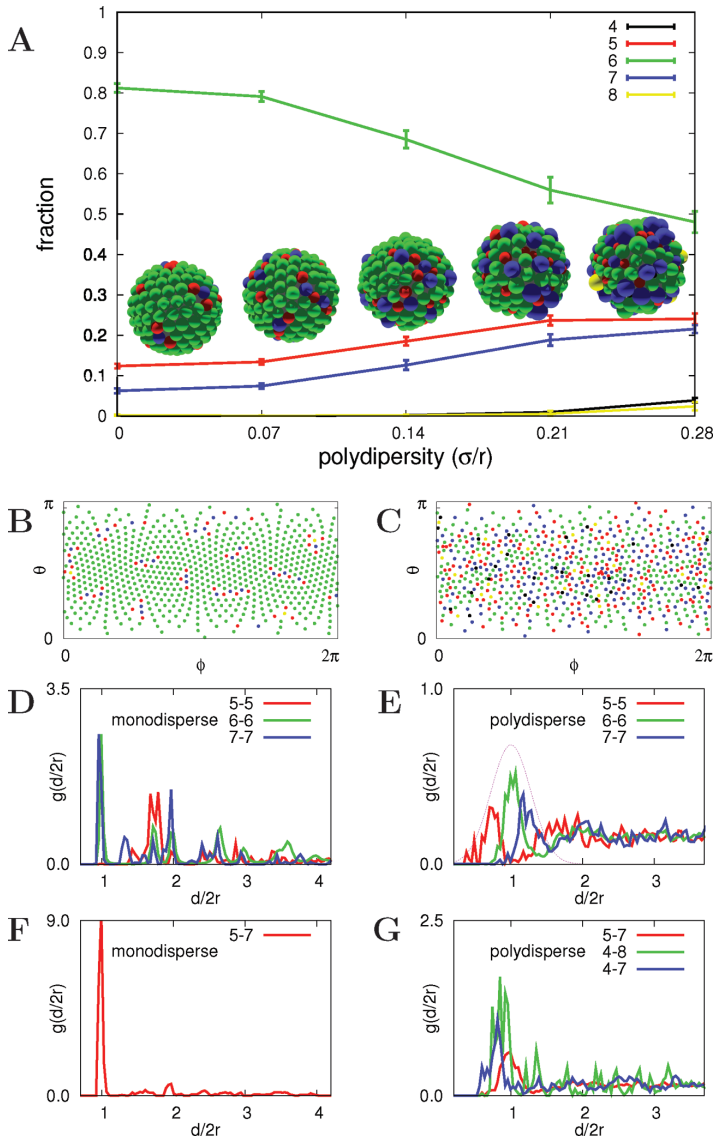


Figure 3.3: (A) Fraction of nanoparticles with different coordination numbers for different values of the polydispersity for $R/r = 6.7$. Equilibrium structures for four different values of polydispersity σ/r are shown. The nanoparticles are colour coded according to the coordination number (3 magenta, 4 black, 5 red, 6 green, 7 blue, 8 yellow, 9 cyan, and 10 brown). (B-C) 2D map of the MC equilibrium structures with $R/r = 13.4$ and $\sigma/r = 0.00$ nm (B) or $\sigma/r = 0.28$ (C) using the same colour code for the coordination numbers used in A. (D-F) Normalised correlation function between spheres with different coordination numbers defined as

$$g(d) = \langle N_{c_a} N_{c_b} \rangle / [A \cdot \langle N_{c_a} \rangle \cdot \langle N_{c_b} \rangle]$$

where $\langle N_{c_a} N_{c_b} \rangle$ is the number of particles with coordination number c_b at a radial distance between d and $d + \Delta d$ from particles with coordination number c_a, $A = 2\pi \sin(d)\Delta d$ is the area of the spherical ring, $\langle {}_{c_a} \rangle$ and $\langle {}_{c_b} \rangle$ are the number of particles with coordination numbers of c_a and c_b, respectively, and r is the average radius of the nanoparticles. The averages have been calculated over 200 configurations.

map shown in Fig. 3.3B for the system with $R/r = 13.4$, and as the polydispersity increases, the scars merge together and extend to cover the whole surface as shown in the map of Fig. 3.3C.

To understand how the surface structure is affected by a change in polydispersity, we cannot simply describe the system in terms of scars or isolated defects, but we can use the correlation function between the position of nanoparticles with a given coordination number as a descriptor. We reported the correlation function between 5-, 6-, and 7-fold spheres in Fig. 3.3D (for the monodisperse case) and Fig. 3.3E for a system with polydispersity $/ra = 0.28$. Fig. 3.3D illustrates the long-range order present in the monodisperse case and the fact that 5-coordinate spheres are never found in contact because the curvature of the droplet is too large to support this energetically unfavourable arrangement. In the polydisperse system (Fig. 3.3E), nanoparticle with coordination numbers of 5, 6, and 7 are found in contact (as expected) at increasing values of the interparticle distance, and the system appears to be more disordered overall. We included in Fig. 3.3E the (Gaussian) distribution of distances between any two particles to illustrate how the broadening of the correlation functions is due to the fluctuation of the nanoparticle size. Fig. 3.3F, showing the correlation function between particles with coordination numbers of 5 and 7, demonstrates how the 5-7 pair substitutes for the 6-6 pairs in the monodisperse case and can be thought to be the dominating defect in the structure (where other hetero-coordinated pairs are very rare). In the polydisperse case (Fig. 3.3G), we observe several hetero-coordinated pairs, and it is not possible to characterise them as defects with respect to a reference structure.

3.4 Conclusion

We have demonstrated that packing patterns of silica nanoparticles on submicrometer droplets of styrene, solidified *via* a Pickering miniemulsion polymerisation process, can be predicted with MC simulations using a 12-24 LJ potential. Furthermore we have shown that by broadening the particle size distribution of the nanoparticles, the packing geometry can no longer be described in terms of a 12-points dislocation or grain-boundary scars.

4 2D HEXAGONAL LATTICE MODEL FOR THE HYDROGEN-BONDED DRIVEN SELF-ASSEMBLY

Simplicity is the ultimate sophistication.

(Leonardo da Vinci, 1452-1519)

We model the 2D self-assembly of planar molecules capable of complementary interactions (like H-bonding) as a set of hexagonal tiles on a hexagonal lattice. We use MC simulations to study the phase diagrams of three model systems. The phases are characterised using a variety of order parameters and they are studied as a function of the strength of the complementary interaction energy. This simplified model is proven to be capable of reproducing the phases encountered in real systems, unifying within the same framework most of the structures encountered experimentally.

4.1 2D Self-Assembly

The H-bond driven self-assembly of molecular systems adsorbed onto a surface has recently attracted great interest due to the ability of the building blocks to autonomously form nanoscale regular patterns [20, 21, 22]. A variety of systems have been studied, which have been shown to form several 2D structures on surfaces, with common motifs including square [166, 167] and hexagonal networks [19, 168, 169, 170], lamellar [171], and honeycomb structures [172, 173, 24, 174]. As the molecular building blocks are often based on aromatic benzene rings (such as benzene-dicarboxylic acids[19, 168, 169]), the underlying symmetry of the molecules leads to hexagonal networks, with each molecule having six neighbours.

In almost all the cases, the molecules interact with their neighbours through H-bonds and Van der Waals interactions (although other interactions, such as π-stacking [170], have also been observed), and, in Fig. 4.1, we show how it is possible to group the building blocks forming hexagonal networks on the basis

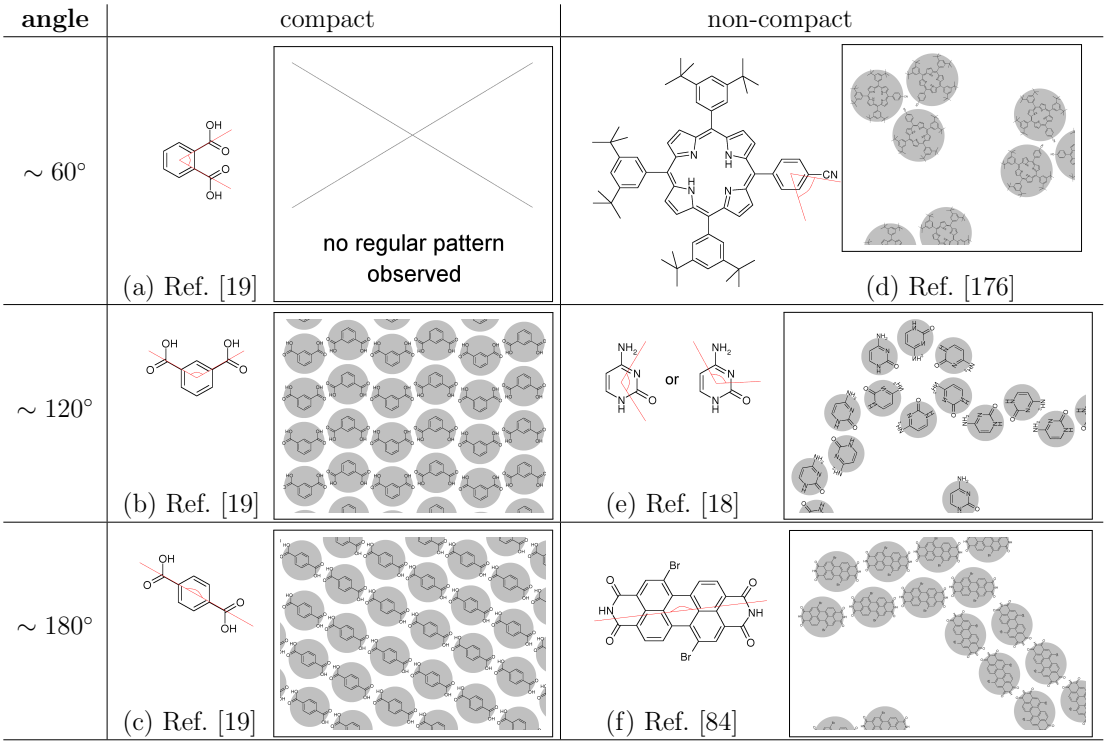


Figure 4.1: Schematic representation of the 2D structures formed by a representative set of molecules with different angles between the groups with complementary interaction. Compact structures (left) are observed for (b) terephthalic acid on graphite [19], and (c) isophthalic acid on graphite [19]. Non-compact structures (right) are observed for (d) CTBPP on Au(111) [176], (e) cytosine on Au(111) [18], and (f) Br-PTCDI on Ag–Si(111) [84].

of the angular distance between the strongest binding directions. Molecules with a 120 or 180 degrees between the H-bonds, such as meta or para-dicarboxylic acids [19, 168, 169], are capable of forming regular structures (Fig. 4.1b-c), but no regular patterns have been observed if the angular distance between the complementary groups is equal to 60 degrees (Fig. 4.1a), as happens with phthalic acid [19]. Interestingly some systems, even when presenting the same bonding motif, form non-compact structures in which the order is kept only for a few molecular units. Examples includes 4-Phenylazopyridine [175] or CTBPP [176] on Au(111) which self-assemble into trimers (Fig. 4.1d), cytosine on Au(111) [18] which forms chains and rings (Fig. 4.1e), and Br_2-PTCDI on Ag-Si(111) [84] that forms fibrillar structures (Fig. 4.1f).

Systems that self-assemble on metallic surfaces have been studied theoretically using a number of methods, such as density functional theory (DFT) or MD and

MC. Due to the computational cost, DFT has been resticted to static calculations on small numbers of molecules. DFT calculations have been used to study single molecules [168, 84] and the supramolecular structures formed by small numbers of molecules[177, 84, 85, 86], considered to be the building blocks of larger scale assemblies.

For instance, the deviation from planarity of a set of dicarboxylic acids is associated with a low adsorption energy of the molecule on the surface [168], and the 2D optimisation of molecular chains leads to the conclusion that the H-bonds of a 2D system are extended with respect to those formed in a 3D system [177]. In other studies [84, 85, 86], dimers, trimers, and tetramers of the same molecule have been investigated to identify the lowest energy supramolecular building block, for instance melamine dimers and tetramers have been studied to identify the structures compatible with the geometry of the Au(111) surface [86]. Mixed *ab initio* and MM methods have been used to study CTBPP on Au(111), using benzonitrile as a model [178], to confirm the experimental structure [176] (Fig. 4.1d).

In general, the effect of the surface on the self-assembled structure have been studied in more detail with MD [179] and MC [180] simulations with all atom representations; in particular these have been used to identify the absorption sites on the surface. For instance, MD simulations have been used to follow the the motion of DNA nucleobases on a Au(111) surface in order to calculate the adsorption energies and the mobility of the molecules [179] on the surface. Kinetic MC simulated annealing has been used to study oxalic amide derivatives assemblies which appear to be templated by the Ag(111) surface properties [180], and to study the patterns formed by PTCDA and melamine on Au(111) [181] performing 2D simulations (assuming no effect of the surface other than keeping the structure planar). Phase diagrams have been calculated for mixtures; for example a thermodynamic equilibrium model has been employed to describe the phases observed experimentally for a mixture of 1,3,5-benzenetribenzoic acid (BTB) and 1,3,5-tricarboxybenzene trimesic acid (TMA) [182], and a lattice model for a melamine-PTCDI mixture [183].

As the available theoretical studies are in general concerned with the interpretation of specific experimental systems, a general understanding of the relationship between the shape and structure of the building blocks and the supramolecular structures they can form on surfaces is still lacking. This work aims at building a common framework able to rationalise the broad range of structures that one can form on a planar surface. It has been largely shown that it is possible to construct simplified models that can reveal how the building blocks structure can affect the

phase behaviour of a system, as observed in real systems [184]. Such simplified models have been used to study a wide range of systems, including liquid crystals [185], polymers, and proteins[186]. In this context, lattice models, in which only the essential features of real systems are retained, have proven to be particularly useful to describe the rich phase behaviour that can be observed experimentally. For example, lattice models have a long history in the study of polymers [56], for the behaviour of proteins and other biomacromolecules [57, 58], liquid crystals [59, 60], and the most stable self-assembled structures that a set of particles can form with an AB technique [43, 187].

Here, due to the particular symmetry of the problem, we will model molecular systems forming hexagonal networks as a set of tiles on a hexagonal lattice. Hexagonal lattices have been extensively used for the study of the magnetic properties of materials, with Ising models [52] and Potts models [50]. Ising models can give us useful information about the phase transitions of particles systems [188]. Unfortunately, molecular systems are far more complicated and more detailed models should be developed to study their properties. A step in this direction is the "poker chip model", used to study the directional ordering of proteins in a 2D hexagonal lattice [51] by Higo and collaborators. In their model, each protein is represented as a hexagonal chip. The chips are free to perform discrete rotatations but their position on the lattice is fixed to the lattice points. The interaction between two neighbouring chips is described by assigning a coefficient to each edge of the chips, and multiplying the coefficients at the contact point to take into account the relative orientations of the two proteins, which affects the interaction strength between the two. The critical behaviour of this model has been studied by Hogyoku [189]. Similar models with continuous degrees of freedom have been used to study the behaviour of patchy colloidal particles [186, 190] as models of proteins. A more recent example of hexagonal lattice model has been proposed by Weber and collaborators, to simulate the self-assembly of a melamine-PTCDI molecular mixture deposited on a metal surface by gas deposition [183] using MC in the NVT ensemble. Their model, where the molecules occupy the edges and vertexes of a hexagonal lattice, successfully simulates the self-assembled structure formed by gas deposition.

Here simulations are carried out in the NVT ensemble, in order to simulate the self-assembled structures formed by planar molecules capable of complementary interactions, such as H-bonding; unlike Ref. [183] the molecules of our model occupy the hexagonal sites of the lattice, instead of its edges. Moreover, we will consider systems composed of identical molecules and discuss three different classes of molecular systems. In the next section, our hexagonal-lattice model for the H-bond

driven self-assembly is introduced, then the structures and phases expressed by the system are identified and analysed. Finally, the simulation results are discussed in the context of the available experimental data.

4.2 Model system and Order Parameters

We consider a set of identical non overlapping hexagonal tiles on a 2D hexagonal lattice with periodic boundary conditions (PBC). Adjacent tiles interact with a Van der Waals-like interaction plus an additional electrostatic-like contribution, if two charged edges are in contact. Each edge l of a tile i has a formal "charge" $Q_i^{(l)}$ and the Hamiltonian of the system is:

$$H = \frac{1}{2}\sum_{\langle ii'\rangle}(-\epsilon) + \frac{1}{2}\sum_{\langle il,i'l'\rangle} Q_i^{(l)} Q_{i'}^{(l')} \tag{4.1}$$

where the summation $\langle ii'\rangle$ extends over all adjacent tiles i and i' and the summation $\langle il, i'l'\rangle$ extends only over the contact edges l and l' of the adjacent tiles i and i'. We consider the three charge distributions depicted in Fig. 4.2: in the tiles of type A two consecutive edges are charged with opposite polarity, in the tiles of type B two nonconsecutive edges are charged with opposite polarity, and in the tiles of type C the opposite edges are charged. The absolute value of the charge on the charged edges is set to the parameter q. The charged edges can be thought of as donor-acceptor H-bonding groups, or any group capable of complementary and local interaction.

The model is easy to generalise to more complicated interactions, but in this minimal form it is already capable of expressing a rich phase behaviour. For example, the tiles A, B, and C can be thought of as representing three benzene-dicarboxylic acids with substituents in ortho, meta, para positions respectively. In the model, when two opposite charged edges are in contact, we will say they are forming a H-bond, and the average number of H-bonds per tile $\bar{\eta}_{\mathrm{H}}$ is a useful measure to characterise the phases expressed by the system [183].

We performed Metropolis MC simulations where the tiles are allowed to move on the lattice. The tiles move through attempted translations, rotations either about the C_6 axes orthoganal to the tile or about one of the six C_2 axes in the plane of the surface. At each simulation step a move is randomly performed and accepted with a Metropolis acceptance probability [83].

On the lattice, each tile i has an associated orientation e^i with respect to the lattice frame (depicted as arrows, in Fig. 4.2). As the tile orientations belong to

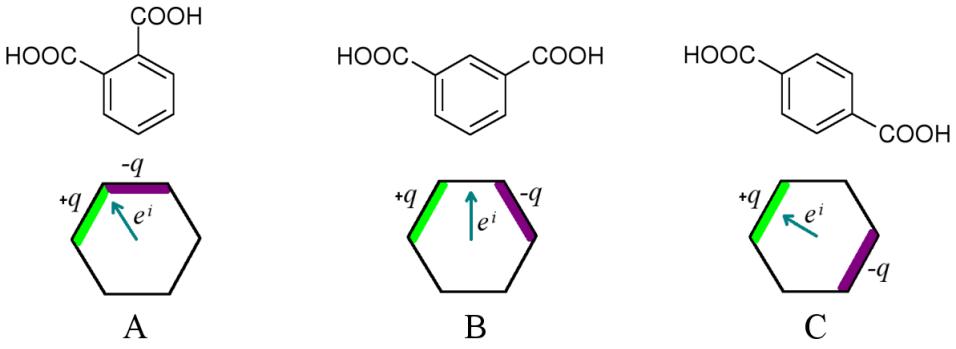


Figure 4.2: The three different hexagonal tiles considered in this work, building blocks of the self-assembled structures. A tile interacts with other tiles through a Wan der Waals like interaction through each edge, plus an electrostatic interaction acting only between the marked edges. The tiles A, B, and C can be thought of as models of the three benzene-dicarboxylic acids with substituents in ortho, meta, para positions respectively. The arrows indicate the vectors e^i relative to the orientation of the tiles with respect to the lattice frame.

a discrete set ($e^i \in \{e_n, -e_n\}$, $n = 1, 2, 3$), we can identify the preferred orientation as the most commonly found orientation [191]. The degree of order may be characterised though the orientational order parameter S. As many of the ordered structures formed by these molecules either involve anti-parallel pairs of molecules or are identical when the symmetry axis of the molecules is flipped 180°, it is convenient to consider an order parameter that is invariant under the transformation from e^i to $-e^i$. Explicitly the order parameter is given by

$$S = \frac{3}{2}\frac{\max_n \left(N_{e_n} + N_{-e_n}\right)}{\sum_i N_{e^i}} - \frac{1}{2} \tag{4.2}$$

where N_{e_n} is the number of tiles with orientation e_n and $\sum_i N_{e^i}$ is the total number of particles in the system. The form of the order parameter defined above is so that it lies in the range $[0, 1]$: $S = 1$ indicates perfect orientational order and $S = 0$ indicates that there is no orientational ordering.

To identify phase transitions, we also calculate the specific heat at constant volume C_V, given by the fluctuation formula:

$$k_{\mathrm{B}} T^2 C_V = \langle (E - \langle E \rangle)^2 \rangle \equiv (\Delta E)^2 \tag{4.3}$$

where E is the total energy of the system at a given simulation step, given by Eq. 4.1, and $\langle E \rangle$ is the average energy calculated over the simulation, after equili-

bration.

Finally, distinct phases can be characterised and differentiated by their interface extension $\bar{\eta}_{\mathrm{I}}$, given by the average number of empty cells in contact with each tile, and the fraction of accepted MC moves. Compact phases, such as liquid and solid phases, tend to minimise their interface, gas phases and porous materials, such as gels, do not. The fraction of accepted MC moves is, for a lattice system, a measure of the diffusivity of the particles in a certain phase: we expect a lattice gas to be free to translate and rotate, a lattice liquid to have frozen translational degrees of freedom, and a lattice solid to have frozen all the degrees of freedom.

To characterise the phase transitions and identify the phase boundaries, we monitored both the maximum of the C_V and the maximum of its first derivative (associated with a first and second order phase transition, respectively). Phase transitions may also be associated with a sudden change in the interface extension $\bar{\eta}_{\mathrm{I}}$, a change in the order parameter S, a change in the number of H-bonds per tile $\bar{\eta}_{\mathrm{H}}$, therefore the derivative of these quantities is also useful to determine the phase boundaries.

4.3 Results

We performed a set of simulations for every tile in Fig. 4.2, where only two edges are charged with opposite polarity $+q$ and $-q$, to investigate their phase behaviour as a function of the bonding parameter q. We investigated values of q from $q = 1$ to $q = 10$. Each simulation was performed at constant density and temperature, using 1250 particles, considering lattice sizes from 36×36 (96% coverage, density $d = 0.96$) to 112×112 (10% coverage, density $d = 0.10$). PBC have been used. Throughout this section, all the energies are expressed in units of ϵ, and the reduced temperature is defined as $T^* = k_{\mathrm{B}}T/q^2$.

For every value of q and for every tile type, a chain of simulations was built. Every simulation of the chain was performed at a lower temperature with respect to the previous one. We explored reduced temperatures from 1.0 to 0.001 for tiles A and C, and from 0.5 to 0.0005 for tile B. Each chain started with $T^* = 1.0$ (tiles A and C), or $T^* = 0.5$ (tile B), and each successive simulation was run lowering the temperature of the previous simulation by $\Delta T^* = 0.001$ (tiles A and C), or $\Delta T^* = 0.0005$ (tile B). We performed a chain of 1,000 simulations, each corresponding to a temperature step. At every temperature step 10^8 MC moves were attempted, and the resulting configuration was then used as starting configuration for the next temperature step. In addition, we looked for hysteresis in the phase diagram: we repeated the simulation chains for each tile starting from

the lowest temperature end configurations and $T^* = 0.001$ (tiles A and C), or $T^* = 0.0005$ (tile B), and increasing the temperature by $\Delta T^* = 0.001$ (tiles A and C), or $\Delta T^* = 0.0005$ (tile B) at every temperature step. No significant differences were found between the two sets of simulations, and therefore, only results from the cooling cycle will be presented.

To analyse the details of the phase diagrams associated with each particle, we calculated the orientational order parameter S, heat capacity C_V, average number of H-bonds per tile $\bar{\eta}_{\mathrm{H}}$, and interface extension $\bar{\eta}_{\mathrm{I}}$ as a function of T^*, for each tile and, for each function, we calculated a ΔX function as $\Delta X = |X(T^*) - X(T^* - \Delta T^*)|$, where $X(T^*)$ are the moving averages of each function calculated over 50 temperature steps.

At high temperature, all the systems show a *gas* phase, in which the tiles are free to move and occupy all the available volume. At intermediate temperatures, some systems show a *liquid* phase, in which the tiles are packed but not yet frozen in their positions/orientations. At low temperatures, the solid structures that we encountered are illustrated in Fig. 4.3 (and quantitatively discussed in the next subsection). Each system presents one of five possible solid structures, depending on the value of q; the possible solid structures are (i) *compact aperiodic*, observed for tile A and small values of q where the tiles form trimers and the trimers pack without forming a periodic structure, (ii) *supramolecular gel*, observed for tile A and large values of q, in which the trimers are arranged into a network that extends to occupy all the available surface, (iii) *crystal*, formed by both tiles B and C, at low values of q, in which the tiles appear ordered and well packed, (iv) *gel*, observed for tile B and large values of q, in which we observe a disordered solid arranged into a network that spans all the available space, and (v) *fibres*, formed by tile C that, at large values of q forms elongated structures.

Gels are disordered solids, whose components are arranged into a network, that usually extends to occupy all the available volume [192, 193]. In the case of tile B, at larger values of q, we observe honeycomb structures and zig-zag chains that extend to occupy the whole simulation box. In the case of tiles A and large q, the tiles associate into trimers and the trimers forms chains which span the whole available space, that is why we call this structure supramolecular gel.

In the rest of this section, first we analyse the effect of the temperature on each system, and the relative phase diagram, discussing a set of simulations on a 50×50 lattice; then we analyse the effect of the density on the phase diagrams.

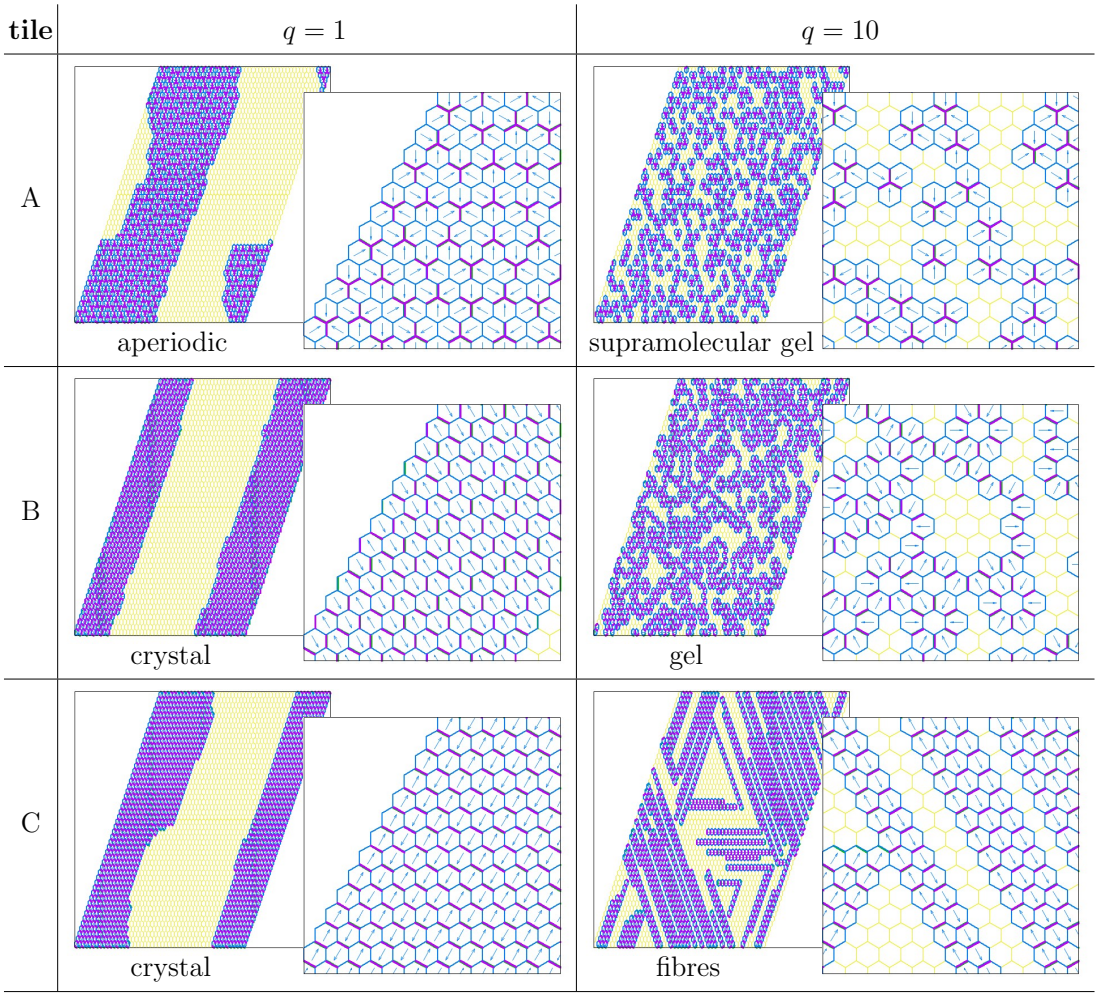


Figure 4.3: The solid phases obtained at low temperature for systems composed of tiles A, B, C at 50% coverage ($d = 0.50$) for $q = 1$ and $q = 10$. A magnification of each structure is also shown.

4.3.1 Effect of the Temperature and Phase Diagrams

Tile A The phase diagram for tile A at 50% coverage is shown in Fig. 4.4. In this system we observe 4 phases: gas, liquid, compact aperiodic (or solid), and supramolecular gel. The phase boundaries are determined from the maximum of $\Delta\bar{\eta}_{\mathrm{H}}$, and the maximum of C_V. Interestingly, the phases expressed by tile A are characterised by a lack of orientational order at every temperature ($S \simeq 0$, Fig. 4.5b).

At high temperatures, at every value of q the system behaves like a gas: in

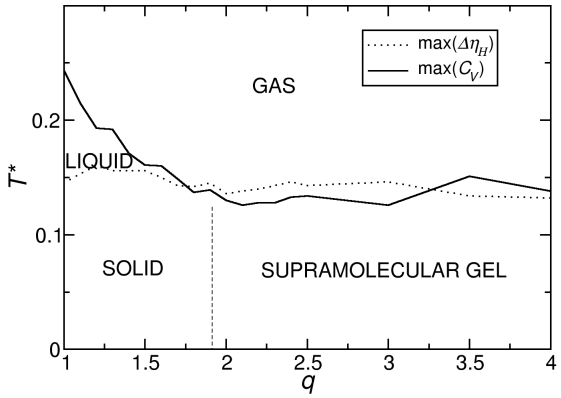


Figure 4.4: Tile A, phase diagram at 50% coverage. The phase boundaries are identified by the maxima of $\Delta\bar{\eta}_{\mathrm{H}}$ (dotted line), and C_V (solid line). The vertical dashed line indicates the value of q beyond which no transition is seen in $\bar{\eta}_{\mathrm{I}}$ (see text).

this phase, each tile is free to translate and rotate (Fig. 4.5e-f), and the interface extension $\bar{\eta}_{\mathrm{I}}$ is equal to 3 (Fig. 4.5d), which is the maximum possible value for $\bar{\eta}_{\mathrm{I}}$, given that the system coverage is equal to 50% (i.e., if the tiles occupy random sites on the lattice, each tile has on average 3 occupied and 3 empty sites in its neighbourhood).

As the temperature decreases, the phases encountered by the system depend on the value of q. At low values of q the system crosses two phase transitions, the highest temperature one characterised by a peak in C_V and the lowest temperature one by a shoulder in the same parameter (Fig. 4.5a): the first phase transition has an associated change in the interface extension $\bar{\eta}_{\mathrm{I}}$ (Fig. 4.5d), the second phase transition is instead associated to a change in the number of H-bonds per tile $\bar{\eta}_{\mathrm{H}}$ (Fig. 4.5c). The phase between these two phase transitions is a liquid phase, characterised by the minimisation of $\bar{\eta}_{\mathrm{I}}$ (Fig. 4.5d), a lack of orientational order (Fig. 4.5b) and frozen translational degrees of freedom (Fig. 4.5f). The phase at low temperature corresponds instead to a packed solid, in which all the motions are frozen (Fig. 4.5e-f). This solid lacks long range order, but presents a local ordering of the tiles, which form trimers that, in turn, are randomly packed (see Fig. 4.3A - left). As q increases the first phase transition shifts towards lower reduced temperatures (Fig. 4.5a) and, at $q = 1.5$, it overlaps with the second phase transition. At $q = 1.8$ the solid phase is not longer packed, as can be assumed by

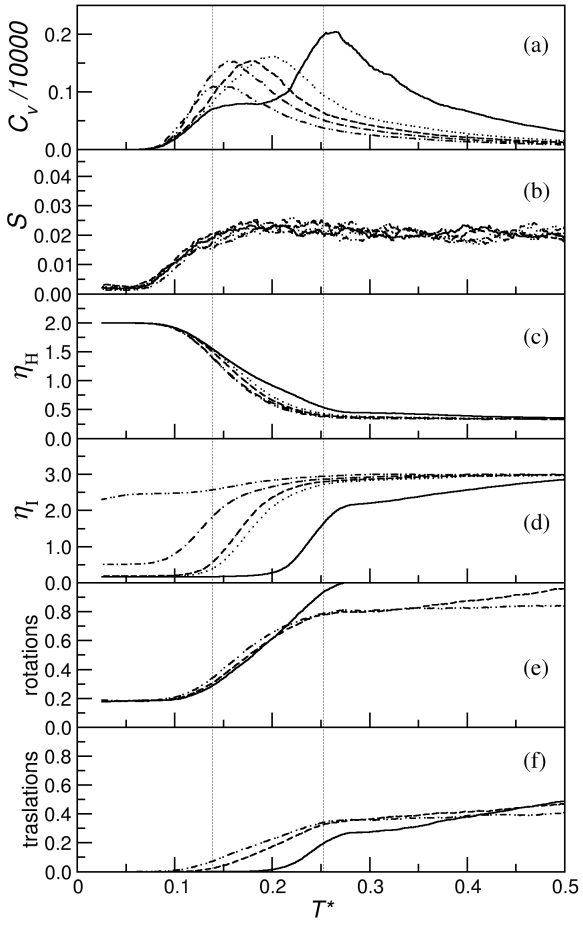


Figure 4.5: Tile A, parameters study at 50% coverage. (a) heat capacity C_V, (b) order parameter S, (c) average number of H-like bonds per tile $\bar{\eta}_{\mathrm{H}}$, (d) interface per tile $\bar{\eta}_{\mathrm{I}}$, (e) accepted rotational and (f) translational moves as a function of the temperature T^* for $q = 1.0$ (solid line), $q = 1.3$ (dotted line), $q = 1.5$ (dashed line), $q = 1.8$ (dashed-dotted line), and $q = 4.0$ (dashed-dotted-dotted line). The dashed vertical lines identify the two phase transition temperatures of the system with $q = 1$.

a smaller decrease of $\bar{\eta}_{\mathrm{I}}$ at low temperature, with respect to that at lower values of q (Fig. 4.5d). This solid presents the same local structure as that of the packed solid, namely it is formed of trimers (see Fig. 4.3A (right)), but the trimers are not closely packed, instead they are associated to form a supramolecular gel.

Tile A, which has the groups with the complementary interactions on two adjacent edges, can therefore pack densely, but cannot form ordered structures. The packing can be controlled by tuning the energy separation between differ-

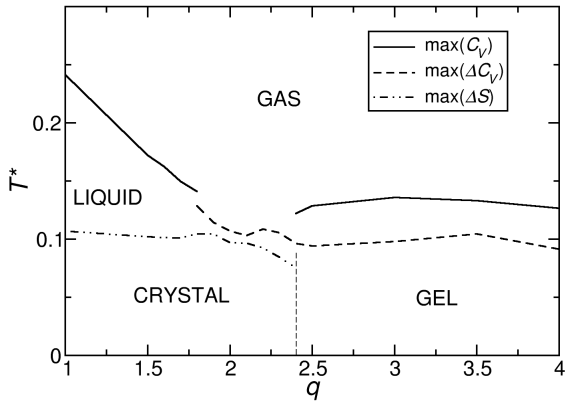


Figure 4.6: Tile B, phase diagram at 50% coverage. The phase boundaries are identified by the maxima of $\Delta\bar{\eta}_{\mathrm{H}}$ (dotted line), C_V (solid line), ΔC_V (dashed line), ΔE (solid-dotted line), ΔS (dashed-dotted-dotted line), and $\Delta\bar{\eta}_{\mathrm{I}}$ (dashed-dotted-dashed line). The vertical dashed line indicates the value of q beyond which no transition is seen in $\bar{\eta}_{\mathrm{I}}$ (see text).

ent sides of the molecule; experimentally this may be controlled by changing the donor/acceptor groups.

Tile B The phase diagram for tile B is shown in Fig. 4.6. Tile B shows gas, liquid, crystalline, and gel phases. The gas-liquid transition and the liquid-crystal transition are analogous to that observed for tile A, as can be verified by comparing Fig. 4.5 and Fig. 4.7 in the case of $q = 1$. The only difference is that the liquid-crystal phase transition in this system can be also identified by the maximum of ΔS, which takes place at the same point as that of the maximum of $\Delta\bar{\eta}_{\mathrm{H}}$ (Fig. 4.7b-c).

As for tile A, at low values of q ($q < 1.7$) the system undergoes two phase transitions upon cooling: first a gas-liquid transition, followed by a freezing transition. At intermediate values of q ($1.7 \leq q \leq 2.4$) we observe a gas-crystal transition identified by the maximum of ΔC_V and the maximum of ΔS. At higher values of q ($q \geq 2.5$) the gas-gel transition, peculiar to this particular bonding pattern, occurs between the maximum of C_V and the maximum of ΔC_V. With respect to the ongoing discussion on where to place the transition temperature in a gel system, given that different system parameters are modified at different temperatures [192], our results suggest that it would be reasonable to define a "transition

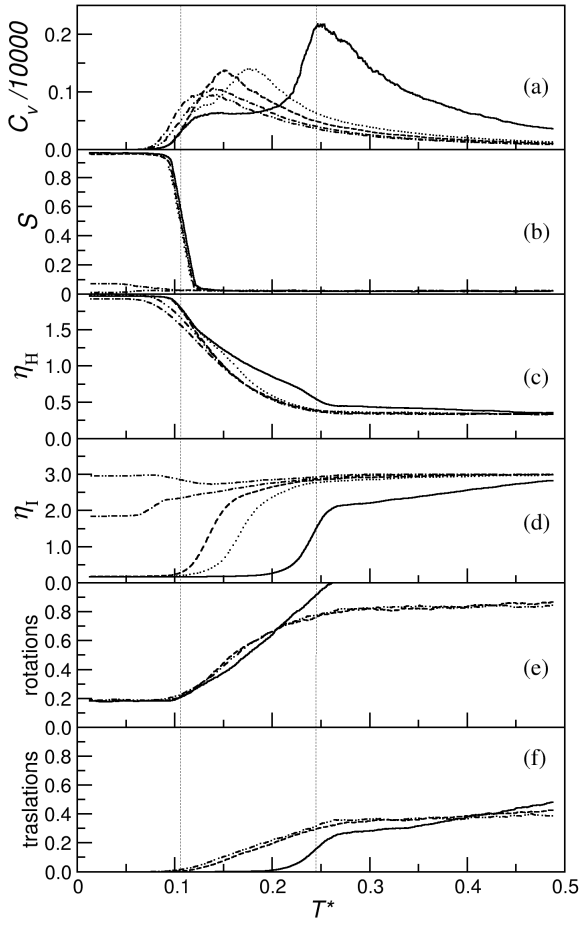


Figure 4.7: Tile B, parameters study at 50% coverage.. (a) heat capacity C_V, (b) order parameter S, (c) average number of H-like bonds per tile $\bar{\eta}_{\mathrm{H}}$, (d) interface per tile $\bar{\eta}_{\mathrm{I}}$, (e) accepted rotational and (f) translational moves as a function of the temperature T^* for $q = 1.0$ (solid line), $q = 1.5$ (dotted line), $q = 1.8$ (dashed line), $q = 2.5$ (dashed-dotted line), and $q = 4.0$ (dashed-dotted-dotted line). The dashed vertical lines identify the two phase transition temperatures of the system with $q = 1$.

interval of temperatures", instead of a single transition temperature. In fact, in this window of temperatures a sequence of events happens: at higher temperature the C_V reaches its maximum value, here we also observe the maximum of $\Delta\bar{\eta}_{\mathrm{H}}$ (Fig. 4.7c) and the maximum of the slope of the energy function (not shown), and at lower temperature the maximum of $\Delta\bar{\eta}_{\mathrm{I}}$ (Fig. 4.7d) curve is found at the same temperature as that of the maximum of the slope of the C_V curve.

An additional feature of this phase diagram is that, as q increases, the gas-liquid

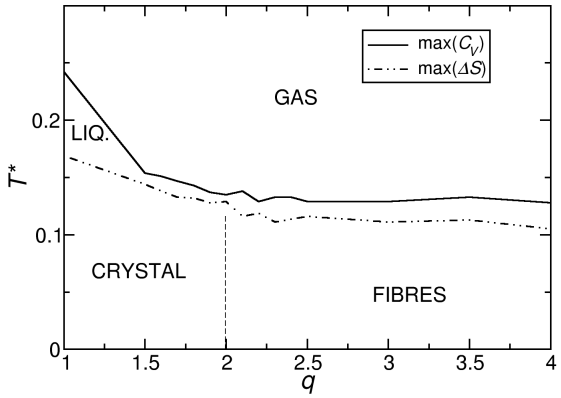


Figure 4.8: Tile C, phase diagram at 50% coverage. The phase boundaries are identified by the maxima of $\Delta\bar{\eta}_{\mathrm{H}}$ (dotted line), C_V (solid line), ΔE (solid-dotted line), and $\Delta\bar{\eta}_{\mathrm{I}}$ (dashed-dotted-dashed line). The maximum of ΔC_V (dashed line) is plotted for completeness. The vertical dashed line indicates the value of q beyond which no transition is seen in $\bar{\eta}_{\mathrm{I}}$ (see text).

phase transition shifts towards lower temperatures (Fig. 4.7a), and, at $q = 1.7$, it overlaps with the liquid-solid phase transitions. At this point we have simultaneous packing and ordering of the system. The low temperature structure we observe is indistinguishable from the crystal structure in Fig. 4.3B. At $q \geq 2.4$, the system is no longer capable, under the chosen simulation conditions, to reach a perfectly ordered state ($S < 1$, at low temperature, Fig. 4.7b), and the single phase transition changes its characteristics: it is no longer associated with a sudden change in S, but it remains associated with a change in $\bar{\eta}_{\mathrm{H}}$. The low temperature structure we observe at $q \geq 2.4$ is the gel structure in Fig. 4.3B (right).

From these result we argue that molecules with the bonding motif of tile B are likely to be trapped in a gel phase for large values of the complementary bonding strength (q). Unlike tile A, this bonding pattern shows an ordered structure when the energy separation between different sides of the molecule is small. Therefore, also in this case, the key to achieve a packed structure is to design the building blocks such that the stronger complementary interactions are no greater than 2.5 times the weaker interactions.

Tile C The phases we observe for tile C are gas, liquid, crystalline and fibres (Fig. 4.8). Unlike the previously discussed systems, this system does not present

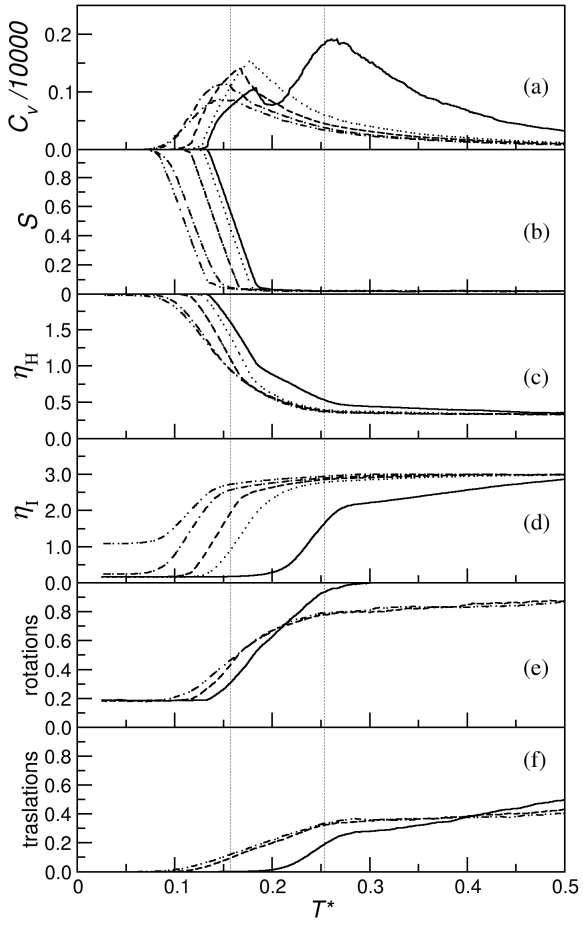


Figure 4.9: Tile C, parameters study at 50% coverage. (a) heat capacity C_V, (b) order parameter S, (c) average number of H-like bonds per tile $\bar{\eta}_{\mathrm{H}}$, (d) interface per tile $\bar{\eta}_{\mathrm{I}}$, (e) accepted rotational and (f) translational moves as a function of the temperature T^* for $q = 1.0$ (solid line), $q = 1.5$ (dotted line), $q = 1.8$ (dashed line), $q = 2.5$ (dashed-dotted line), and $q = 4.0$ (dashed-dotted-dotted line). The dashed vertical lines identify the two phase transition temperatures of the system with $q = 1$.

disordered solid phases. At low temperature the system is aligned at every value of q (Fig. 4.9b). The phase boundaries can be identified by plotting the maxima of ΔS and of C_V. Also in this case, as for tile A, the maximum of ΔC_V is never associated with any other property transition. The gas-liquid and liquid-crystal transitions are analogous to the gas-liquid and liquid-crystal transitions observed for tile B, the first associated with a change in C_V and a drop in η_{I}, and the second associate with the maximum of ΔS and the saturation of the H-bonds (η_{H}, Fig. 4.9c). But the gas-

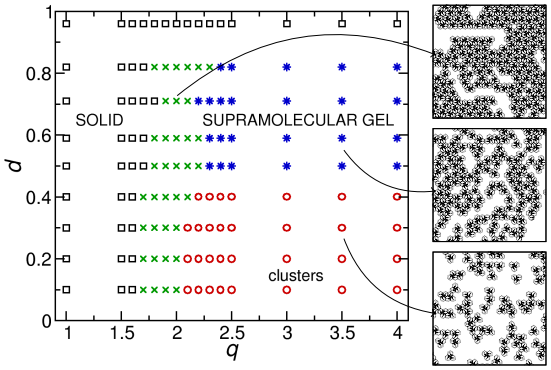


Figure 4.10: Tile A: d vs. q phase diagram, at $T^* = 0.001$. At low temperature we observe a solid (squares), a colloidal gel (stars), isolated clusters (circles), or the coexistence of two phases (crosses), depending on the interaction strength q and the system density.

fibres transition has its own characteristics and it appears to happen in two steps: at higher temperature the H-bonds are formed (Fig. 4.9c) and this corresponds to the maximum of C_V (Fig. 4.9a), at lower temperature the interface extension is minimised (Fig. 4.9d) and this correspond to the maximum of ΔS (Fig. 4.9b). The high temperature part of this phase transition is analogous to that of the gas-gel transition of tile B. The low temperature part of the phase transition is different because there is a variation of the orientational order parameter S, but the change in the second derivative of the C_V which, in this case, appears at even lower temperatures, is not associated with any other property change (not shown). In comparison with tiles A and B, the liquid phase is stable over a smaller region of the phase diagram.

Another feature that differentiates tile C from tiles A and B is that, in this case, both phase transitions shift towards lower temperatures as q increases, as indicated by a shift of both the average number of H-bonds per tile (Fig. 4.9c) and the orientational order parameter S (Fig. 4.9b). At $q > 2$, under the current simulation conditions, the system is not able to minimise the interface extension (Fig. 4.9d).

The H-bond strength can therefore be tuned to switch between a packed solid and fibres (as in Fig. 4.3C). The bonding strength is not, however, the only parameter that needs to be tuned. As we will see in the following subsection, the lattice coverage turns out to be an important controlling parameter in the fibres formation.

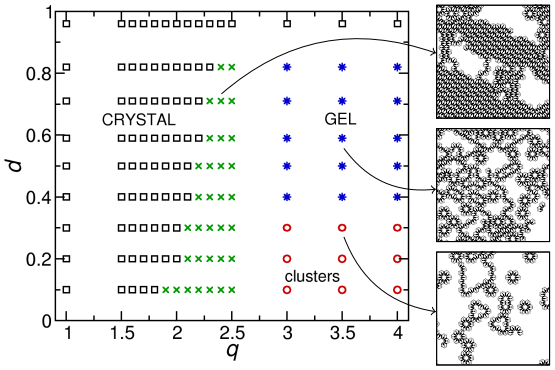


Figure 4.11: Tile B: d vs. q phase diagram, at $T^* = 0.0005$. At low temperature we observe a crystal (squares), a gel (stars), isolated clusters (circles), or the coexistence of two phases (crosses), depending on the interaction strength q and the system density.

4.3.2 Effect of the Lattice Coverage

To study the effect of the lattice coverage, we performed a set of simulations at constant density, using 1250 particles, considering lattice sizes from 36×36 (96% coverage, density $d = 0.96$) to 112×112 (10% coverage, density $d = 0.10$). We rationalise here the end simulation configurations ($T^* = 0.001$ for tiles A and C, and $T^* = 0.0005$ for tile B) in a set of three phase diagrams.

Tile A As previously pointed out, tile A solid structures consist of the assembly of trimers (as in Fig. 4.3A) which can form a packed solid or a supramolecular gel (Fig. 4.10). The density does not have any effect on the packed solid observed at low q ($q \leq 1.6$), but it does affect the non packed structures observed at higher values of q. Due to the limited system size, we are unable to identify accurate phase boundaries, however we can identify the typical behaviour observed in systems that present gel phases [192, 194]: at low densities ($d < 0.5$) we observe isolated clusters and at high densities we observe a gel that spans all the available space. At intermediate values of q we generally observe the coexistence of packed and non-packed structures, which usually appears as isolated trimers coexisting with a wide surface occupied by the solid. The lack of a packed structure over a wide range of q for this bonding pattern is consistent with STM studies on phthalic acid [19] and other molecules [175, 176].

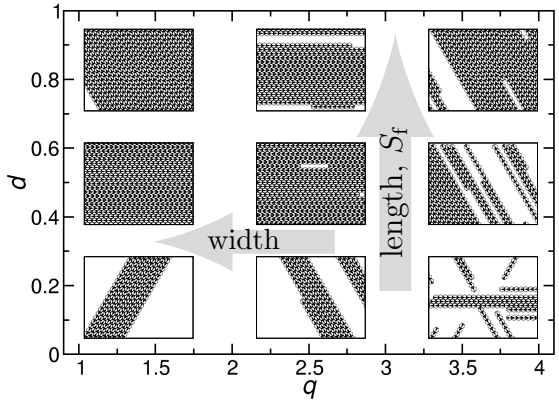


Figure 4.12: Tile C: d vs. q scheme, at $T^* = 0.001$. At low temperature we observe different fibre conformations depending on the interaction strength and the system density. The vertical and horizontal arrows indicate the direction of increasing fibre lengths and of increasing fibre widths, respectively.

Tile B The phase diagram for tile B (Fig. 4.11) is analogous to that of tile A as three phases are observed, with a coexistence region. Differences are that isolated clusters are observed at low densities, where the building blocks assemble into hexamers and chains, and the gel appears at lower densities ($d > 0.3$) with respect to what observed for tile A. The region of coexistence between the crystal and the non-packed structures is pushed towards higher densities as q increases. For example at $q = 2.2$ the two solid phases, disordered gel and ordered crystal, coexist. Thus it is possible to obtain a packed structure through increasing density. This is not the case if $q > 2.3$: here we observe the coexistence of crystal and gel at every density. These results partially explain why a density increase is sometimes insufficient to promote ordered structures, as happens with the cytosine deposition on Au(111)[18] where cytosine molecules are arranged in a disordered structure even when the coverage is increased (see Fig. 4.1e).

Tile C For tile C it is not possible to draw a phase diagram as for the other two systems. The transition between the observed structures occurs smoothly through all the explored range of parameters. In this system (Fig. 4.12) at low q and high d we observe a crystalline structure, at high q and low density we observe short unidimensional fibres. It is possible however to characterise the fibres through their width and length. Moreover, as the fibres are formed by unidimensional fibrils, composed of isoaligned tiles connected through their charged edges, it is possible

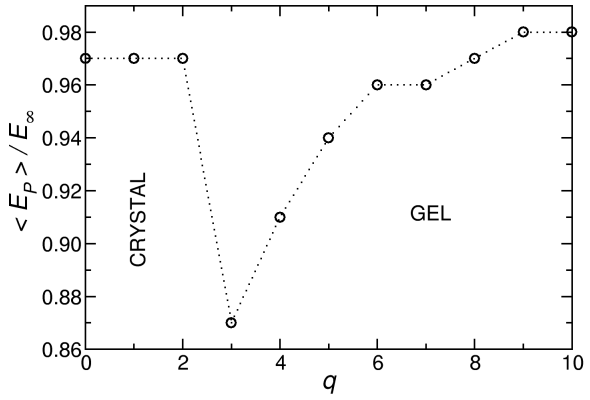


Figure 4.13: Tile B. The ratio $\frac{\langle E_{\rm P}\rangle}{E_\infty}$ is plotted against q.

to assign a direction to each fibril and calculate a fibril order parameter $S_{\rm f}$ as in Eq. 4.2. The fibres elongate and align when the system density is increased and their width increases as the complementary interaction strength decreases.

The results for tile C suggest that fibres morphology is affected by the interplay between complementary interactions and sample density. Moreover, as for tile B, if the desired result is a crystal, an increase in density is not enough to achieve such a structure if the complementary interactions are too strong (e.g. $q > 2$).

4.4 Discussion and Conclusion

In this chapter we discussed a 2D lattice model for the study of H-bonded molecular networks deposited on a surface which brings into a unified framework the different structures that one can encounter with the most common bonding motifs. The model is able to reproduce the experimentally observed crystalline compact structure, and three experimentally observed non-compact structures (gel, colloidal gel, and fibres). It also predicts an aperiodic compact structure not yet reported by the experiments. Moreover, our model shows how changes in the bonding strength of the building blocks can be used to control the transition between gel/colloid/fibres states and compact phases: as the separation between the energies increases, it is more difficult to get the lowest energy configuration, in agreement with the experiments [195].

To understand why these particles do not form a packed structure at zero temperature and every value of q, we focus our attention on tile B of Fig. 4.2, in the

case of $d = 0.50$ and we calculate the average energy per particle ($\langle E_{\mathrm{P}} \rangle = \langle E \rangle / N$) in the simulated system, and, as a reference, the energy per particle (E_∞) for an infinite regular solid (density $d = 1.00$). Given that each tile can interact with 6 neighbours with the Van der Waals like interaction and with 2 tiles with an electrostatic contribution, we get:

$$E_\infty = -(3 + q^2) \tag{4.4}$$

The ratio $\frac{\langle E_{\mathrm{P}} \rangle}{E_\infty}$, plotted in Fig. 4.13, has an interesting behaviour as a function of q: the ratio falls suddenly from ~ 0.97 to 0.87 as q increases and slowly it recovers to 0.98. In all cases the ratio is not unitary because of the effect of the interface. This result suggests that at high values of q, a less packed structure is most likely to appear as it is energetically favourable to form extended structures to maximise the electrostatic interaction, at the cost of decreasing the non-specific interaction due to the van der Waals interactions.

Despite its simplicity, the proposed model can be related to and is consistent with the experimental observations. For example, meta and para-benzene-dicarboxylic acids form regular patterns, while the ortho has never been observed forming regular structures [19]. To connect our model with realistic energy values, it is possible to proceed as follows. Given that, in our model, two charged edges interact with $E_{\mathrm{strong}} = -(\epsilon + q^2)$ and two non-charged edges interact with $E_{\mathrm{weak}} = -\epsilon$, and all the energies are expressed in units of ϵ, the value of q can be calculated as $\sqrt{\left|\frac{E_{\mathrm{strong}} - E_{\mathrm{weak}}}{E_{\mathrm{weak}}}\right|}$, where E_{strong} is referred to the strongest binding energy (e.g. the energy related to the H-bonded configuration) and E_{weak} to the weaker (e.g. the van der Waals energy of two molecules non H-bonded). The values for E_{strong} and E_{weak} can be calculated with MM methods or taken from the literature. For example, calculations using MM3 force-field (described in Appendix A.3) show that the meta benzene dicarboxylic acid, which has the bonding pattern of tile B, has $E_{\mathrm{weak}} = -1.72$ kcal/mol and $E_{\mathrm{strong}} = -11.16$ kcal/mol which corresponds to $q = 2.34$, and forms a regular structure, in agreement with the experiments [19]. Cytosine interaction energies can instead be found in the literature [196]. Using the estimates of Ref.[196] ($E_{\mathrm{weak}} \simeq 0.05 E_{\mathrm{strong}}$ with $E_{\mathrm{strong}} = -20.7$ kcal/mol) we get $q \simeq 4.36$. The cytosine also presents the same bonding pattern as tile B, and its value of q corresponds to the formation of a 2D gel, which is in agreement with the experiment [18].

This model can be therefore considered intermediate between a purely phenomenological (qualitative) model and a quantitative (atomistic) description of

the patterns formed by H-bonded molecules. As the full characterisation of the phase diagram is too computational expensive for an atomistic study we suggest that a simulation approach like the one proposed here is the most suitable to rationalise the vast amount of experimental information available.

5 AGENT BASED ALGORITHM FOR THE STUDY OF MOLECULAR SELF-ORGANISATION

Adapt or perish, now as ever, is nature's inexorable imperative.

(H. G. Wells - A Short History of the World, 1922)

N THIS CHAPTER, an algorithm based on the concept of AB simulations is developed to predict the lowest energy packing of a set of identical rigid molecules. The agents are identified with rigid portions of the system under investigation, and they evolve following a set of rules designed to drive the system towards the lowest energy minimum. The algorithm is compared with a conventional Metropolis MC algorithm, and it is applied to a large set of representative models of molecules. For all the systems studied, the AB method consistently finds a significantly lower energy minimum than the MC algorithm because the system evolution includes elements of adaptation (new configurations induce new types of moves) and learning (past successful choices are repeated).

5.1 Predicting Self-Assembled Structures

Although considerable progress has been made in the prediction of self-assembled structures, it is still generally not possible to predict the most stable structure formed by an aggregate of identical molecules [197]. The difficulty of the problem is also testified by the fact that the great effort put into the prediction of the crystal structure of organic molecules [38] or the self-assembled structure of simple compounds [198] often leads to results that are unsatisfactory.

Standard simulation methods as MD or MC are unable to correctly predict the most stable aggregates because the self-assembly process spans multiple time and length scales, giving rise to a complicated free energy surface with many local minima [7], which act as kinetic traps for both types of simulation. In the case of MD, many strategies such as hyperdynamics [90], metadynamics [92], or temperature accelerated methods [91] may be used to sample more rapidly the phase

space, but these methods are ill suited to simulate particle aggregation into an *a priori* unknown structure. MC-based methods offer, in principle, the possibility of performing "unphysical" moves in order to escape from kinetic traps, but these moves are, again, difficult to design if the nature of the kinetic traps and the desired optimal structure are not known [199, 200, 201]. An important step forward in the development of MC algorithms for self-assembly was the introduction of cluster moves [107, 202, 106] that allow the displacement of portions of the system of variable size and mixed MC-stochastic dynamics methods, which alternate MC and stochastic dynamics steps [184].

An alternative approach to the problem can be found considering that most self-assembling systems of interest are composed of a collection of identical particles, which are effectively searching for the same optimal energy minimum. Exploiting this symmetry of the system, one can devise an algorithm that uses all identical particles to perform a parallel search of the global minimum. This idea is explored in this chapter where we present a new AB algorithm [130] for finding potential energy minima for off-lattice systems (of identical particles). AB algorithms have been developed and used to study complex systems in research areas very far from chemistry (computer science, economics, social sciences [203, 204, 124, 205, 206]).

The main feature of AB models is that they try to describe the collective behaviour of a set of agents. An agent is defined as a computer system capable of sensing its environment, taking decisions and performing autonomous actions (the possible actions and the conditions under which they are undertaken are called the rules of the agent). In AB simulations, the rules of the agent are defined by the user and the outcome of the simulation is the global behaviour of a set of agents interacting with the environment and amongst themselves. Rule-based models are encountered in the modelling of many physical systems [122] (especially in the form of CA models [207, 208, 209, 126, 210, 211, 212]) and constitute a useful alternative to modelling in terms of differential equations. The idea of using AB models for molecular simulation, first proposed by Mark Ratner, was initially implemented for lattice systems [43], where an agent was identified with a shape or a group of shapes on a square lattice. The model proposed in Ref. [43] was able to identify stable aggregates much more efficiently than a standard MC algorithm because the rules incorporated elements of learning and adaptation: the agent/particle learns from other agents and exploits their "experience" to better explore the configurational space. AB algorithms are one of the many realisations of AI based on population learning. A more common AI approach for the global optimisation of chemical systems is given by the GA [113, 118]. In both the GA and AB methods, the

system evolution does not follow the physical evolution of the system, but, while AB models are particularly designed for the study of self-assembly of identical particles, GAs are generally used to optimise clusters or large individual molecules.

The essential component for the development of an AB model is the design of an efficient set of rules that allow the evolution of the system towards the lowest possible energy (possibly coinciding with the optimal self-organised structure). The AB algorithm presented in this chapter will be able to simulate the self-assembly of rigid particles off-lattice. Although some of the ideas of this algorithm are also present in Ref.[43], the off-lattice simulation requires, as we will see, the definition of additional rules. The agent, identified with a stable subportion of the system (one or more particles), will evolve during the simulation by optimising its structure, merging with other agents, eliminating particles, moving randomly in space, and leaving disassembled aggregates. In summary, the changes with respect to the original lattice AB algorithm [43] are: (i) reformulation of the merge and split rules, to adapt them to possible moves performed in an off-lattice system; and (ii) inclusion of the disaggregate rule, to separate agent aggregates. Moreover, the presented algorithm is designed to make it usable in conjunction with standard MM computational packages for future applications on realistic chemical systems (as in Chap. 6).

In the next section the rules followed by the agent are defined and justified. In Sec. 5.3 the model systems are introduced and the performance of the AB algorithm is critically evaluated. A discussion (Sec. 5.4) concludes the chapter.

5.2 The Algorithm

5.2.1 Definitions

The system is composed of N rigid particles in a fixed volume with PBC. We describe the configuration and the moves of the system at two distinct levels: (i) The particles, characterised by their position and orientation in space, are allowed to perform single particle MC moves (i.e., translations and rotations); (ii) The agents, identified by a set of particles, may perform a more complicated set of actions (namely move, merge, split, and disaggregate, which are described below). We define the composition A_i of an agent i as the set of particles $\{i_1, \cdots, i_{n_i}\}$ forming a cluster, where n_i is the dimension of A_i, that is, the number of particles in the agent. The starting dimension of A_i is 1 (i.e., each agent initially contains only one particle and all particles are assigned to an agent). The agent can change its composition as an effect of the merge or split action. We define the internal

energy E_i of the ith agent as:

$$E_i = \sum_{i_k=1}^{n_i} \sum_{i_l=i_k+1}^{n_i} V_{i_k i_l} \qquad \text{with} \qquad i_k, i_l \in A_i \tag{5.1}$$

where $V_{i_k i_l}$ is a pair potential between the particles i_k and i_l belonging to the ith agent, and the interaction energy E_{ij} between the two agents i and j as:

$$E_i j = \sum_{i_k}^{n_i} \sum_{j_l}^{n_j} V_{i_k j_l} \qquad \text{with} \qquad i_k \in A_i, j_l \in A_j \tag{5.2}$$

where $V_{i_k j_l}$ is a pair potential between the particles i_k belonging to the ith agent, and j_l belonging to the jth agent.

Finally, we define $E_{\min}[n]$ as the internal energy of the most stable agent of each dimension n formed during the simulation. As described later, by updating $E_{\min}[n]$ during the simulation, an element of learning is incorporated into the algorithm.

5.2.2 Rules

The possible actions that an agent can perform are *move*, *merge*, *split*, and *disaggregate*. These, together with the conditions under which they are performed, form the set of rules that define completely the AB model. They are described in detail below (and graphically in Fig. 5.1).

Move Both agent moves and single particle moves can be performed. The move rule consists of a conventional Metropolis MC move. A rotation or translation is attempted and the energy of the new configuration is calculated. The new configuration is accepted with the Metropolis acceptance rule (as in Eq. 2.25), where T is the absolute "temperature" of the system. As the rest of the moves do not satisfy detailed balance what we defined here as temperature is really a pseudotemperature that corresponds to the thermodynamic temperature only when the remaining actions are suppressed.

Merge If the interaction energy E_{ij} between the set of particles A_i and A_j associated with the agents i and j is smaller than a threshold E_{M}, then the agents will merge into a new agent m with the associate particle set A_m whose elements will be the union of the elements of the sets A_i and A_j ($A_m \equiv A_i \cup A_j$). In other words, whenever the interaction energy between two agents i and j is sufficiently strong (i.e., $E_{ij} < E_{\mathrm{M}}$), they will evolve in the configurational space as a single

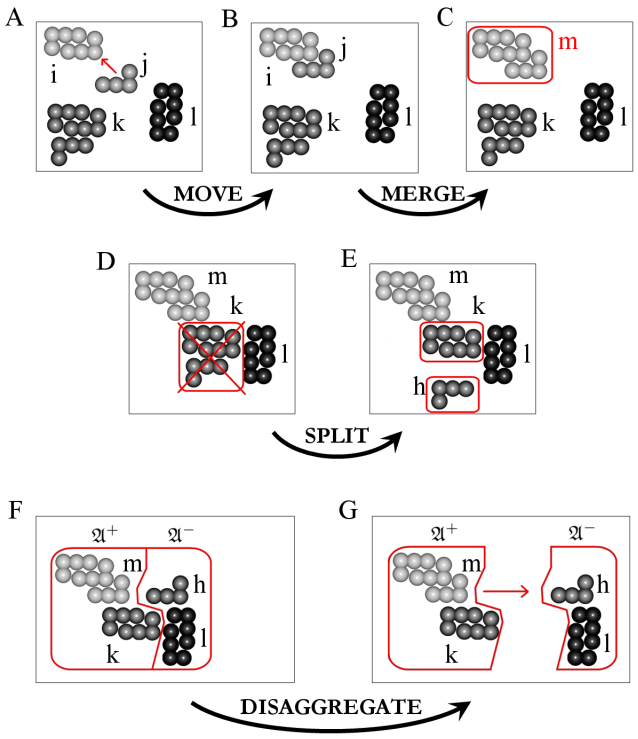


Figure 5.1: Illustration of the basic actions that can be undertaken by the agents (sets of particles). The agent j can perform a simple move in space (from A to B). The agents i and j can merge into a new agent m if their interaction energy is favourable (from B to C) and evolve in space as a unique body. If the energy of an agent k is too high compared with that of other agents in the simulation, then it splits, emitting a new agent h (from D to E). If a set of agents m, k, l, h form an aggregate (kinetic trap) they disaggregate (from F to G). The conditions for each move and their implementation are detailed in the text.

body (when the agent m is moved), exploring the space more effectively. Explicitly, E_{M} is given by:

$$E_{\mathrm{M}} = k_{\mathrm{M}}\{E_{\min}[n_i + n_j] - E_{\min}[n_i] - E_{\min}[n_j]\} \tag{5.3}$$

The quantity in the curly bracket is the difference in energy between the two best clusters of dimension n_i and n_j, and the best cluster of dimension $n_i + n_j$, and it is therefore a good reference energy to establish if the merge action is convenient. The parameter $k_{\mathrm{M}} \in [0, 1]$ defines how often the merge takes place (a low value of k_{M} implies a more frequent clustering of "sub-optimal" structures). In addition, it is convenient to include a further condition in order to avoid the formation of large agents of dimension comparable to N (a single agent much larger than the

others cannot compare its energy with other agents and would not contribute to the parallel search of the optimal minimum). The agents will not merge if the sum of their dimensions is greater than $n_{\max}$. In this work (see Sec. 5.3), the reasonable value $n_{\max} = N/10$ was chosen without specific optimisation.

Split The aim of this action is to destroy suboptimal agents that may be obtained after applying the merge action. This is important because, as the simulation evolves, lower energy clusters may form and the old clusters may no longer fulfil the minimum energy condition defined by $E_{\min}[n]$. If the ith agent (of dimension n_i) has an internal energy E_i which is greater than the energy $k_{\mathrm{S}}E_{\min}[n]$ associated with the most stable agent of the same dimension n_i formed during the simulation, then the agent will split into smaller agents. This simply means that the agent is split whenever a more stable agent of the same dimension has been formed. To keep this rule symmetric to the merge rule, in this work we set $k_{\mathrm{S}} = k_{\mathrm{M}}$.

The decision on which elements of the set A_i to dissociate is based on energy considerations: the particle most weakly bound with the other particles of the ith agent will form a new agent and the energy of the residual agent, with dimension $n_i - 1$ will be compared with $k_{\mathrm{S}}E_{\min}[n_i - 1]$. This is repeated m times until the energy of the residual agent with $n_i - m$ elements is smaller than $k_{\mathrm{S}}E_{\min}[n_i - m]$. This iterative procedure will then produce an agent containing $n_i - m$ elements and m agents containing one element. The m agents are moved far away from the residual agent by performing $1000m$ move actions with pseudotemperature set to 10 times higher than the system temperature.

Disaggregate During the simulation, agents may find themselves in an aggregate of agents which will not merge (because of unfavourable interaction energies). If the number of agents contained in one aggregate is too high, then the agents will not perform any action; they may not move because this would increase the energy and they may not satisfy the criteria for the merge or split action. These aggregates are in a kinetic trap similar to that encountered in conventional MC simulations of self-assembly. The aim of the disaggregate action is to destroy the "wrong" aggregates that are formed during the simulation. Such aggregates may be formed by agents whose attractive energy is too weak to allow the merge action to be performed. Two agents i and j are considered part of the same aggregate $\mathfrak{A}$

if their interaction energy is lower than a certain threshold energy E_{D}:

$$E_{\mathrm{D}} = k_{\mathrm{D}}\{E_{\min}[n_i + n_j] - E_{\min}[n_i] - E_{\min}[n_j]\} \tag{5.4}$$

This definition is formally the same as Eq. 5.3, which defines the merge rule; the difference resides in the different prefactor k_{D} ($k_{\mathrm{D}} < k_{\mathrm{M}}$ always, as we are considering weaker interactions).

As we do not want to compete with the moves that bring the agents together, it is not suitable to split the aggregate as in a split action. Instead, it was found convenient to break the aggregate of agents by partitioning it in two independent sets of agents $\mathfrak{A}^-$ and $\mathfrak{A}^+$. In order to fragment the aggregate into two, while causing the smallest possible increase in energy, we used a technique based on the spectral bisection of a graph. A graph was built where each agent is a node and two nodes are connected if their (negative) interaction energy, divided by the sum of their dimensions, is less than the average interaction energy per particle between the agents in the aggregate. Such a graph can be efficiently partitioned in two (by breaking the smallest number of connections) using a spectral bisection algorithm (see Appendix A.2). After partitioning $\mathfrak{A}$, $\mathfrak{A}^+$ is moved away from $\mathfrak{A}^-$ by 1000 translational moves of the rigid $\mathfrak{A}^+$ with pseudotemperature set 10 times higher than the original temperature.

Initialisation and Update of $E_{\min}[n]$ As stated before, $E_{\min}[n]$ is the internal energy of the most stable agent of dimension n formed during the simulation. The internal energy of the agent is re-evaluated each time that agent changes (particle moves, split, merge). In particular $E_{\min}[n]$ defines the clustering energy E_{M} (see Eq. 5.3) and, every time two agents merge, the internal energy of the new agent should be smaller than the sum of the internal energies of the original agents. To make sure that $E_{\min}[n]$ is properly initialised also when the simulation has not yet explored agents of dimension n we initially set $E_{\min}[n] = (n-1)E_{\min}[2]$. The initial value of $E_{\min}[2]$ is evaluated during the equilibration steps and chosen as the lowest interaction energy between two particles. (For consistency, when $E_{\min}[2]$ is updated during the simulation, the vector elements $E_{\min}[n]$ will be updated to $(n-1)E^{\mathrm{new}}_{\min}[2]$ if $E_{\min}[n] > (n-1)E^{\mathrm{new}}_{\min}[2]$.)

5.2.3 Overall Algorithm

The algorithm is shown schematically in Fig. 5.2. After a number of equilibration steps, which consist of single particles moves, the minimum interaction energy

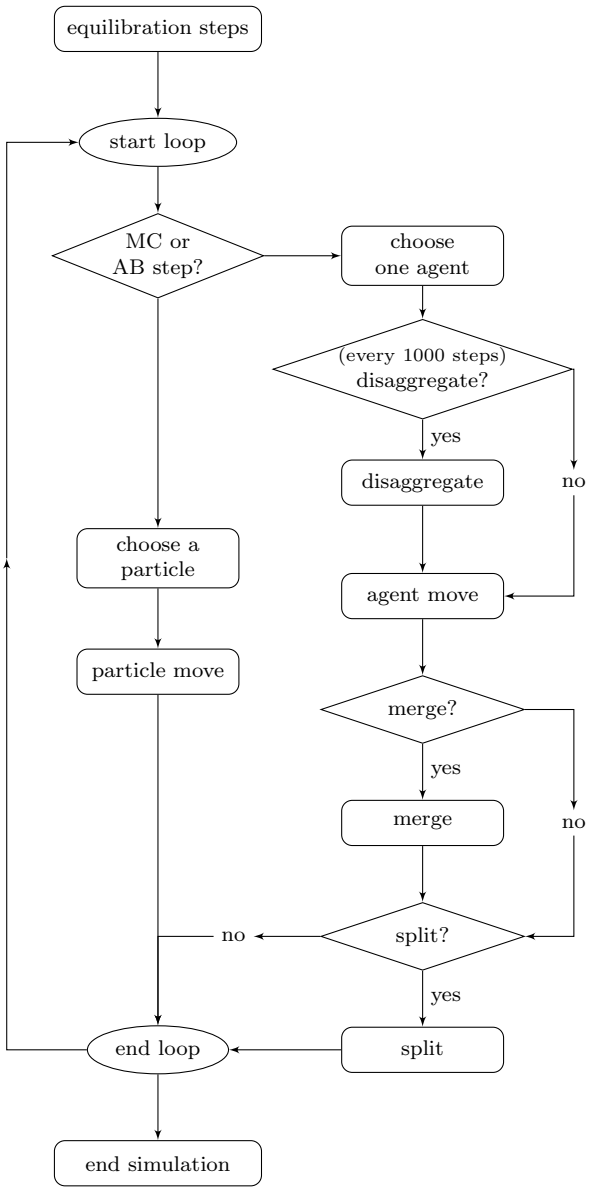


Figure 5.2: Flowchart of the AB algorithm described in the text.

between two particles is evaluated, the vector $E_{\min}[n]$ is initialised, and the main loop starts with the choice of a random number, if this number is lower than a tunable parameter (χ_{MC}, the fraction of MC moves) a single particle step (standard MC) is performed. If the random number is greater than MC, an AB step is performed. The AB step consists of choosing an agent and appling the set of rules previously introduced: a cluster move is attempted and the new configuration is

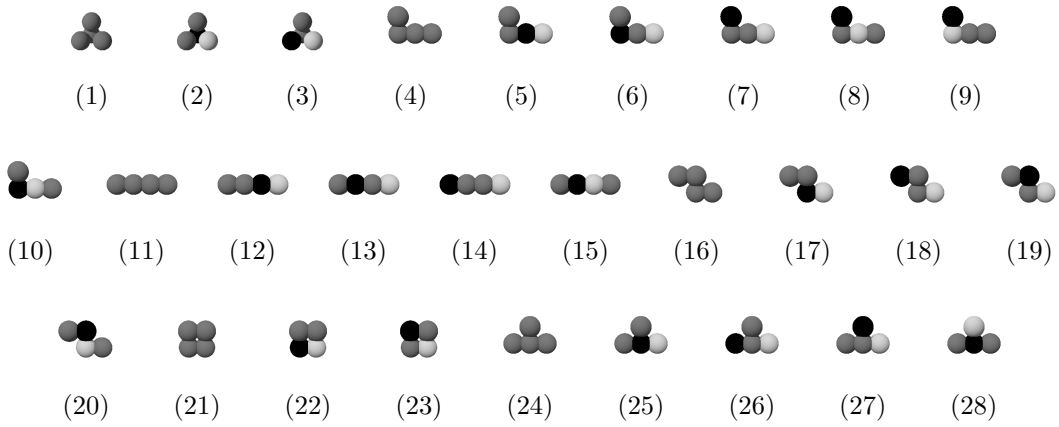


Figure 5.3: The set of particles used to test and develop the algorithm. The greyscale represents the charge of the atoms composing each particle: gray (neutral), white (positive), and black (negative).

accepted or rejected; then, if the interaction energy between the chosen agent and one or more agents is smaller than E_{M}, the agents will merge; finally, the agent energy is compared with E_{S} and, if the internal energy of the agent is greater than E_{S}, the agent will split; otherwise, $E_{\mathrm{min}}[n]$ will be updated. Every 1000 steps the disaggregate rule is applied: the agent is said to form an aggregate with its neighbourhood if all the (negative) interaction energies between the chosen agent and the agents in the neighbourhood are smaller than E_{D}; if the aggregate contains more than four agents, this will also disaggregate.

5.3 Results

5.3.1 Model Systems

The proposed algorithm has been developed, tested, and optimised on a set of 28 rigid particles (illustrated in Fig. 5.3). These were built to constitute an unbiased and representative set that could sample the possible types of kinetic traps during the self-assembly. Each particle is composed of four "atoms" and four possible atomic arrangements are considered (tetrahedral, L shape, I shape, Z shape, square, and T shape). The atoms composing the particles are of three types "neutral" ($Z = 0$), "positive" ($Z = 1$), and "negative" ($Z = -1$) and, for each arrangement, we considered the particle made by all neutral atoms and all symmetry-independent particles containing one positive and one negative atom (with the rest of the atoms

neutral). All the atoms composing the particles interact with a LJ-like potential of the form:

$$V_{ab}^{\mathrm{LJ}} = \left(\frac{1}{r_{ab}}\right)^{12} - 2\left(\frac{1}{r_{ab}}\right)^{6} \tag{5.5}$$

where r_{ab} is the distance between 2 atoms a and b. Eq.5.5 defines the dimensionless units of length and energy used in this work. The non-neutral atoms also interact through an additional Coulomb-like potential

$$V_{ab}^{\mathrm{C}} = \frac{Z_a Z_b}{r_{ab}} \tag{5.6}$$

The total interaction energy between two particles k and l is given by:

$$V_{kl} = \sum_{a}\sum_{b}\left[V_{ab}^{\mathrm{LJ}} + V_{ab}^{\mathrm{C}}\right] \tag{5.7}$$

where the summation is taken over all the atoms a that belong to the particle k and all the atoms b that belong to the particle l. We did not identify these particles with any specific chemical systems but several recent studies have adopted analogous models (rigid particles with pairwise spherical potentials) to study the self-assembly of complex structures [61, 213, 214].

All the simulations were performed in a box of size $30 \times 30 \times 30$ with PBC and the minimum image convention was used, and no cutoff was applied to the Coulomb potential. In all the simulations the system has been first pre-equilibrated for 100,000 MC steps. Following this, production runs of 300,000 steps were used to analyse the effect of the parameters in the preliminary tests (Sec. 5.3.2) and longer runs of 2,000,000 steps were used to study the self-assembled structures (Sec. 5.3.3). In the production phase a step consists of either a MC (single particle) move or one iteration of the AB algorithm. In all the simulations we set $k_{\mathrm{B}}T = 0.10$.

5.3.2 Preliminary Tests

The aim of the preliminary study presented in this section is to tune the AB simulation parameters in order to reach the lowest possible energy configuration in a given number of simulation steps. Starting with a dilute system of particles, we compared systematically the total energy decrease of the AB algorithm (as a function of the number of iterations and the CPU time) with the decrease in energy of a Metropolis MC simulation (taken as a reference algorithm).

Initial simulations were performed by considering the evolution of 50 identical particles, for two types of particles, one with only LJ interactions (particle **4**,

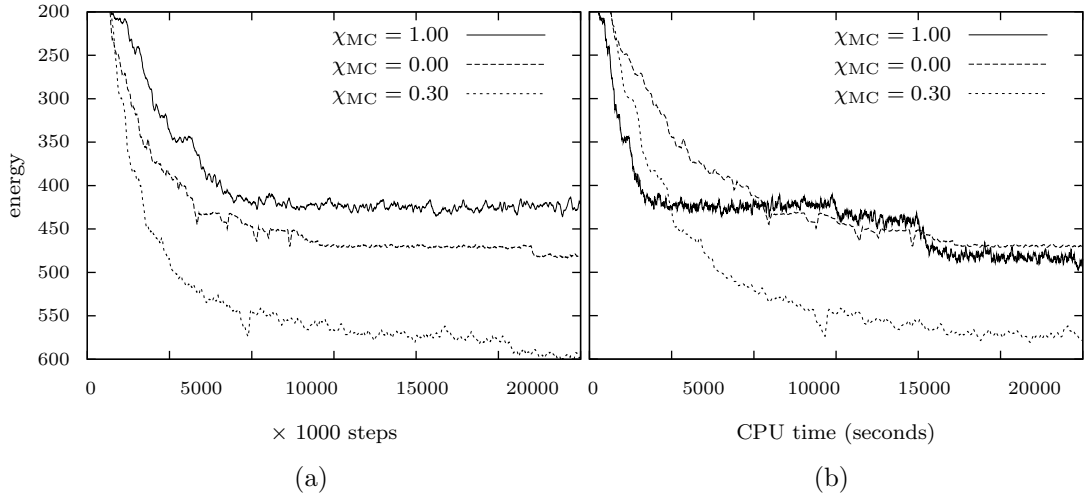


Figure 5.4: System energy vs simulation step (a) and vs CPU time (b) for particle **4** in Fig. 5.3. To make the graph clearer, a moving average of 10 steps was plotted.

Fig. 5.3) and one including the Coulomb potential (particle **26**, Fig. 5.3). The parameter optimisation led to similar results, and we therefore discuss only the results for particle **4**. The parameters studied are χ_{MC} in the range 0.00-0.50, k_M in the range 0.70-0.98, and k_D in the range 0.10-0.50. We considered a grid of 150 parameter sets ($6\chi_{MC} \times 5k_M \times 5k_D$). The collected data are reported in Suppl. S.3.1. Without the inclusion of single particle moves (i.e., fixing $\chi_{MC} = 0$) the set of parameters that give the lowest energy structure is $k_M = 0.98$, $k_D = 0.20$. When single particle moves are included ($\chi_{MC} \neq 0$), the lowest energy structure is achieved with $\chi_{MC} = 0.30$, $k_M = 0.80$, and $k_D = 0.10$. When $\chi_{MC} = 0.30$ the AB algorithm was always able to reach a lower energy configuration (for all k_M and k_D) than a single particle MC algorithm, in the same number of moves.

It is clear from Fig. 5.4a, which shows the performance of the algorithm when applied to particle **4**, that the AB rules have the effect of decreasing the total energy of the system in a smaller number of steps: 400,000 AB steps reach the same or lower energy than 2,000,000 steps of a standard MC simulation. In Fig.5.4b the drop of the energy of the system composed of 50 particles **4** is plotted as a function of the CPU time (measured on a machine with a Intel Xeon 2.66 GHz CPU). For a standard MC simulation, the energy relaxes quickly to a certain configuration with the energy remaining largely constant as the simulation time is increased. The AB algorithm, with the parameters $\chi_{MC} = 0.00$, $k_M = 0.98$, $k_D = 0.20$ is slower than

the MC algorithm because the cost of AB steps is greater than MC steps, in fact, to move an agent i we need to calculate $n_i(N-n_i)$ interaction energies instead of the usual $N-1$ for a MC move. It appears that combining the two techniques with the inclusion of single particle moves in the AB scheme produces an efficient method to reach a low energy structure in a lower number of steps and less computational effort with respect to a standard MC simulation.

The two sets of parameters $\chi_{\mathrm{MC}} = 0.00$, $k_{\mathrm{M}} = 0.98$, $k_{\mathrm{D}} = 0.20$, and $\chi_{\mathrm{MC}} = 0.30$, $k_{\mathrm{M}} = 0.80$, $k_{\mathrm{D}} = 0.10$ have then been applied to the extended set of particles shown in Fig.5.3.

5.3.3 Simulation Results

Three sets of simulations on 50 particles and 150 particles have been compared to measure the actual performances of the AB algorithm: (i) a standard MC simulation ($\chi_{\mathrm{MC}} = 1.00$), (ii) an AB simulation with only agent moves ($\chi_{\mathrm{MC}} = 0.00$, $k_{\mathrm{M}} = 0.98$, $k_{\mathrm{D}} = 0.20$), and (iii) an AB simulation with single particles moves ($\chi_{\mathrm{MC}} = 0.30$, $k_{\mathrm{M}} = 0.80$, $k_{\mathrm{D}} = 0.10$) for all the particles of Fig. 5.3. In this section, the performance of the algorithm has been measured by comparing the average energy of the "equilibrium" configurations defining the relative improvement $\langle R \rangle$ as:

$$\langle R \rangle = \frac{\langle E_{\mathrm{AB}} \rangle - \langle E_{\mathrm{MC}} \rangle}{\langle E_{\mathrm{MC}} \rangle} \tag{5.8}$$

where $\langle E_{\mathrm{MC}} \rangle$ is the average energy of a MC simulation and $\langle E_{\mathrm{AB}} \rangle$ is the average energy of the AB simulation after equilibration (the MC simulation would not have reached thermodynamic equilibrium but a pseudoequilibrium due to the kinetic trap for the formation of the self-assembly). This calculation has been performed over the last 200,000 configurations on 5 runs of the 50-particle system and on the last 1,000,000 configurations on 1 run of the 150-particle system; in both cases we collect one sample every 1,000 steps. The simulation with 150 particles is performed in the same box as the simulation with 50 particles (the system is therefore more dense). The results with 50 or 150 particles are similar and here we will discuss mainly the results with 150 particles. The results for 50 particles are reported in Suppl. S.3.2.

Fig. 5.5, in which $\langle R \rangle$ is plotted against each particle, collects all the results of this test for 50 and 150 particles and for $\chi_{\mathrm{MC}} = 0.00$ and $\chi_{\mathrm{MC}} = 0.30$. This figure shows that the relative improvement for the AB simulations (averaged over the 28 distinct simulations) without single particle moves is 23% for the 50-particle systems and 14% for the 150 particle ones, with two AB simulations not able to

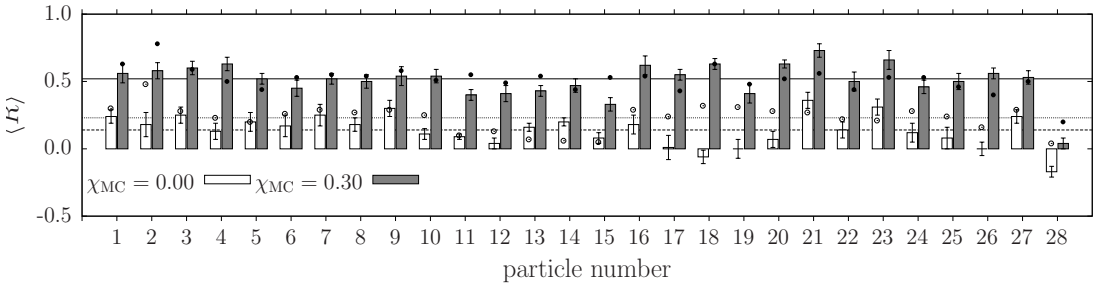
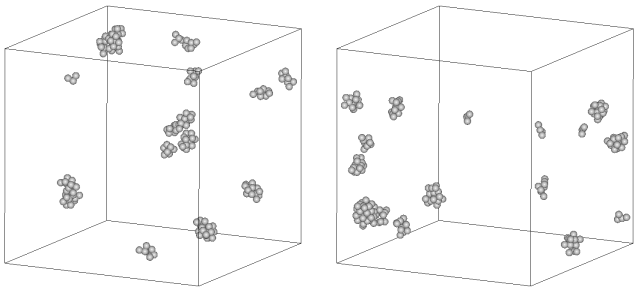


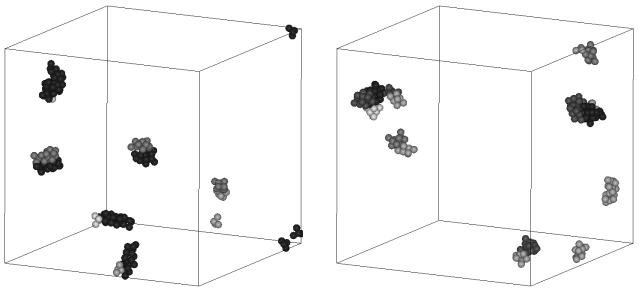
Figure 5.5: Comparison of the relative improvement $\langle R \rangle$ (defined in Eq. 5.8) for the simulation of each of the 28 particles of Fig. 5.3. Here, the histogram represents the averages taken over 150-particle systems: empty bars are used in the case of $\chi_{\rm MC} = 0.00$, $k_{\rm M} = 0.98$, $k_{\rm D} = 0.20$, filled bars for the case of $\chi_{\rm MC} = 0.30$, $k_{\rm M} = 0.80$, $k_{\rm D} = 0.10$. An open (filled) circle on each bar indicates the analogous result for the 50-particle system in the case $\chi_{\rm MC} = 0.00$ ($\chi_{\rm MC} = 0.30$). The dashed, dotted, and solid lines indicate the average relative improvements calculated over all the simulations in the case of $\chi_{\rm MC} = 0.00$ with 50 particles ($\langle R \rangle = 0.23$), $\chi_{\rm MC} = 0.00$ with 150 particles ($\langle R \rangle = 0.14$), and $\chi_{\rm MC} = 0.30$ with both 50 and 150 particles ($\langle R \rangle = 0.52$ and 0.51, respectively).

reach a lower energy configuration. With the inclusion of single particle moves the relative improvement of the AB algorithm is on average equal to the 52% for the 50-particle system and 51% for the 150 particles one, and all the particles reach a lower energy configuration with respect to a standard MC simulation. Consistent with the result of Ref. [43], the AB algorithm without single particle moves performs better with a more dilute system ($\langle R \rangle$ drops from the 23% to the 14%, as we triple the system density). With the inclusion of multiscale moves (i.e., $\chi_{\rm MC} > 0$), the average relative improvement does not change with the system size, allowing the application of this algorithm to denser systems.

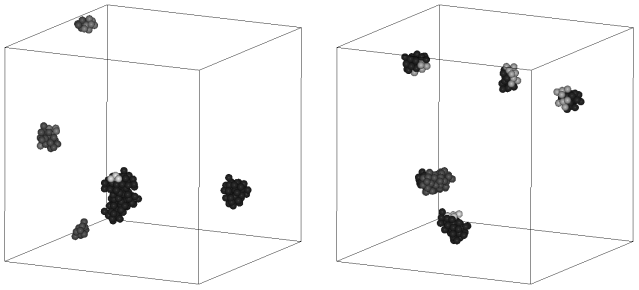
Fig. 5.6 shows the final configuration of the simulation of 50 particles of two types: one that performs better than average (particle type **2**) and one that performs worse than average (particle type **28**) with respect to the total energy minimisation. In both cases the top snapshot corresponds to the MC simulation end-configuration, the second to the AB simulation without single particle moves and the third to a simulation with the inclusion of single particle moves. It is clear from Fig. 5.6 that the AB simulation with the inclusion of single particle moves, third snapshot of each set, gives a more compact structure and larger agents, underlining again the benefits of including multiscale moves (i.e., both single particle



(a) $\chi_{\mathrm{MC}} = 1.00$



(b) $\chi_{\mathrm{MC}} = 0.00$

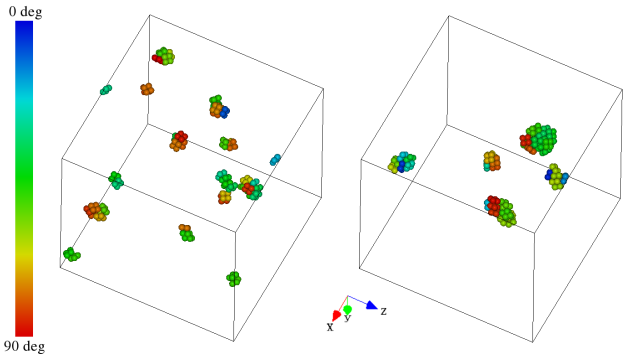


(c) $\chi_{\mathrm{MC}} = 0.30$

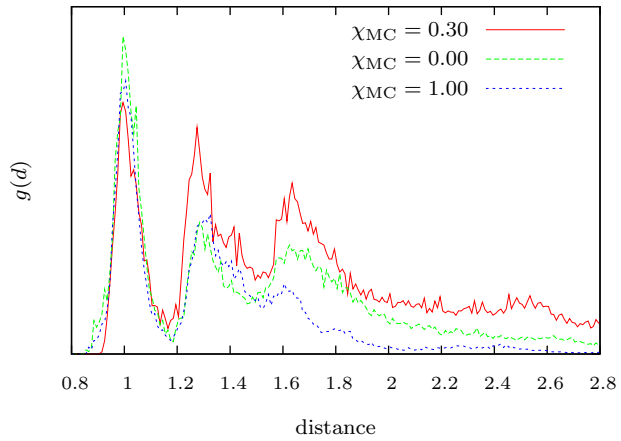
Figure 5.6: Final configuration for the set of simulations performed for particle **2** (left) and for particle **28** (right), reported in Fig. 5.3, for (a) the standard MC simulation; (b) the AB simulation without single particle moves and parameters $\chi_{\mathrm{MC}} = 0.00$, $k_{\mathrm{M}} = 0.98$, $k_{\mathrm{D}} = 0.20$; and (c) the AB simulation with the inclusion of single particle moves, with parameters $\chi_{\mathrm{MC}} = 0.30$, $k_{\mathrm{M}} = 0.80$, and $k_{\mathrm{D}} = 0.10$. The colour code indicates the dimension of the agents (larger agents are darker).

and agent moves) in the model.

The final configurations from the AB simulations are more ordered than those from MC simulations. For example, particle **21** forms only dimers in the MC simulation but, in the AB simulation, the same particle aggregates into more ordered



(a) 50 particles, MC (b) 50 particles, AB



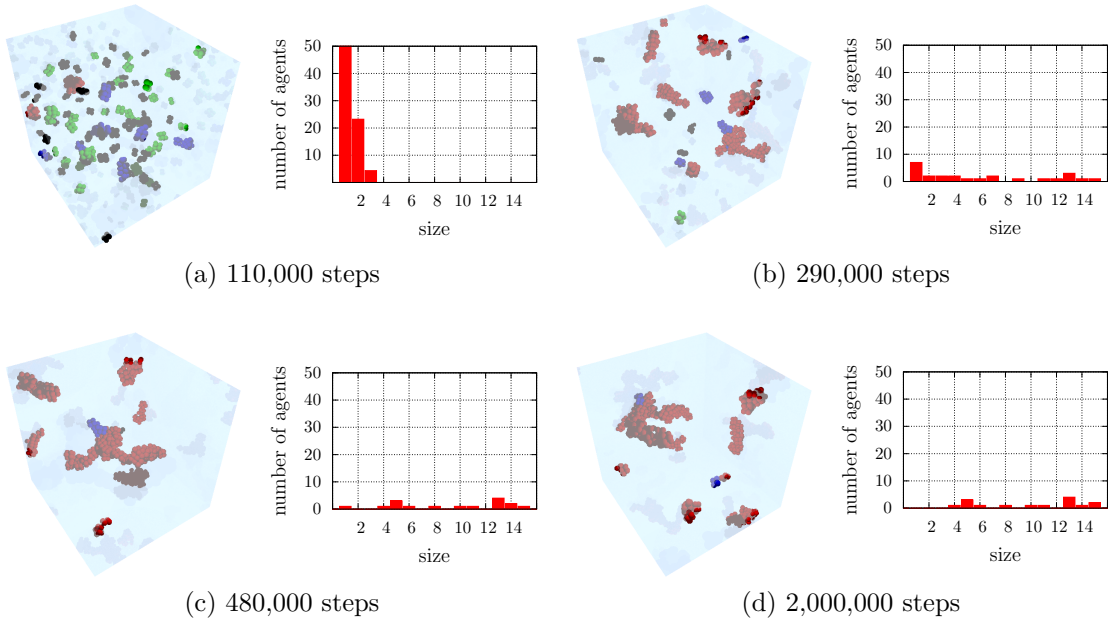
(c) 150 particles

Figure 5.7: (a-b) Final configuration for 50 particles of type **21** for (a) the MC simulation and (b) the AB simulation with parameters $\chi_{MC} = 0.30$, $k_M = 0.80$, $k_D = 0.10$; the colour scale indicates the orientation of the local z-axis with respect to the orientation of the global z-axis, shown in the center. (c) Radial distribution function calculation for the final configuration for 150 particle of the same type. In both representations, the AB simulation gives larger ordered clusters.

clusters as one can easily see in the 50-particle system of Fig. 5.7a-b. In particular, the clusters formed from the AB simulations (Fig. 5.7b) tend to consist of iso-oriented particles. For larger systems, the increased order can be best appreciated plotting the radial distribution function $g(d)$:

$$g(d) = \frac{\langle N^{\text{pairs}}(d) \rangle}{V(d)N} \tag{5.9}$$

where $N^{\text{pairs}}(d)$ is the number of molecular pairs with a certain separation distance



(a) 110,000 steps (b) 290,000 steps

(c) 480,000 steps (d) 2,000,000 steps

Figure 5.8: Agent size evolution for the simulation of particle **21** (same as in Fig. 5.4) with parameters $\chi_{MC} = 0.30$, $k_M = 0.80$, $k_D = 0.10$. (Left) snapshots at different simulation steps with the colour code indicating the agents of size 1 (black), 2 (green), 3 and 4 (blue), and larger than 4 (red). (Right) cluster size distribution for the corresponding simulation step.

d, $V(d)$ is the spherical volume enclosed by the sphere of radius d and the one of radius $d + \Delta d$ (i.e., $V(d) = (4/3)\pi((d + \Delta d)^3 - d^3)$), and N is the total number of particles in the system. The radial distribution function (Fig. 5.7c) shows only one well-resolved peak for the MC simulation where mainly dimers form and no larger ordered aggregate is usually observed; in fact, the peaks relative to trimer and tetramer formation appear as a unique peak. On the other hand, the radial distribution function for the AB simulation shows a sequence of equidistant peaks, indicating that the obtained aggregates are ordered clusters.

It is interesting to see how the agents evolve as the simulation progresses. Fig. 5.8 illustrates the evolution of the system with 150 particles **21**, by comparing the system snapshots with the agent size distribution. In the first 250,000 steps, the monomers rapidly disappear (i.e., from panel a to b in Fig. 5.8) and dimers, trimers, and larger agents form. After about 500,000 steps the agent size distribution does not change dramatically (i.e., from panel c to d in Fig. 5.8 there are minor changes) even if the system energy is decreasing further (as seen in Fig. 5.4), at this stage

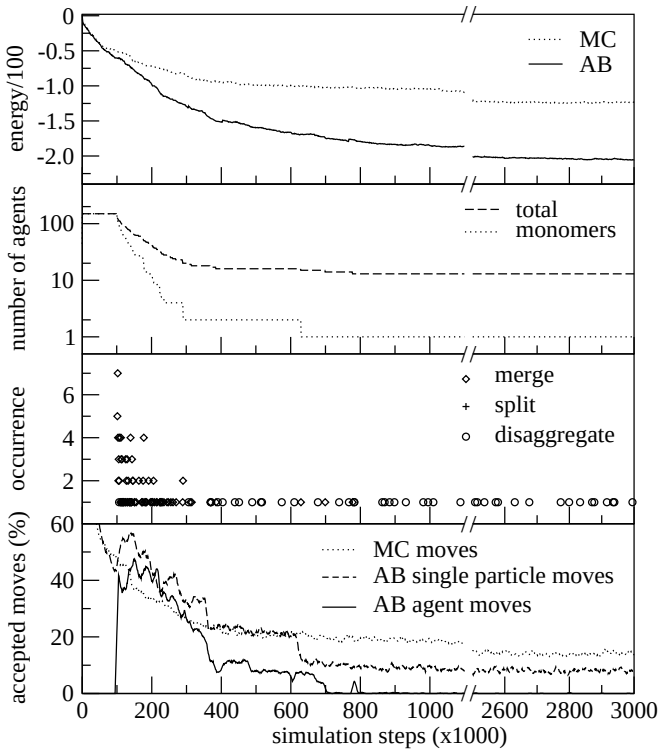


Figure 5.9: Evolution of several simulation parameters for a system containing 150 particles of type **4**. From top to bottom: (i) system energy for the AB simulation (solid line) and a MC simulation (dotted line); (ii) total number of agents (dashed line) and number of agents of dimension 1 (dotted line) at each step of the AB simulation; (iii) occurrence of the actions merge (square), split (cross), and disaggregate (circle) during the AB simulation; (iv) fraction of accepted single particle moves for the MC simulation (dotted line) and the AB simulation (dashed line), and fraction of agent moves (solid line) for the same AB simulation. All the lines are moving averages over 10 data points.

the agents mostly rearrange themselves. It should be noted that the plots of the agent size distribution have been useful in the development of the current version of the algorithm; in particular, the introduction of the disaggregate move, absent in the AB lattice simulations, was made necessary as, without it, larger agents were not formed.

To further illustrate what happens during the simulation and why one can identify two distinct stages of it, in Fig. 5.9 we report a typical simulation run plot (with the parameters reported in Tab. 5.1) in which the occurrence of each action is plotted against the simulation steps. Here it is possible to verify that in the first

description	symbol	value
fraction of single particle moves	χ_{MC}	0.30
Boltzmann x temperature	$k_{\mathrm{B}}T$	0.10
merge definition	k_{M}	0.80
split definition	k_{S}	0.80
disaggregate definition	k_{D}	0.10
maximum agent size	n_{max}	$N/10$

Table 5.1: Simulation parameters and their optimised value.

stage of the simulation, when the agents grow and the monomers disappear, the merge and disaggregate actions are performed very often, whereas in the second stage the move and disaggregate actions are the driving events that lead to further minimisation of the total energy of the system. The split action is unlikely to happen at a temperature in which we expect a solid phase, if the equilibration has been long enough and the $E_{\mathrm{min}}[n]$ array therefore properly initialised. A similar analysis of the simulation evolution can be used to further optimise the simulation parameters for a particular system to be investigated. It is, in principle, also possible to modify the AB simulation parameters as the simulation evolves and to make themselves adaptive. However, in this work, we were interested in studying the general principles for setting up AB simulations, so the AB simulation parameters were held fixed throughout the simulations.

The overall algorithm can be also seen as an accelerated nucleation process. In classical nucleation theory, the rate of nucleation depends linearly on the density of nucleation sites, on the frequency of attempts to form a larger core, and on the probability of successful attempts, the latter depending exponentially on the free energy nucleation barrier [215]. The group moves increase locally the density of nucleation sites (agents tend to move towards each other due to the attractive interactions), whereas the disaggregate rule multiplies the attempts to form a larger core, destroying more quickly the wrong ones. The combination of the two therefore increases the nucleation rate and helps to cross the nucleation barrier.

As an additional test, the results of the AB simulations have also been compared with the results of a cluster MC algorithm, which includes group moves. Several such algorithms have been devised [202, 106, 216] and here we adopt a version developed in our research group based on the concept of DA [217, 108], whose details are given in Appendix A.1. Comparison with a MC algorithm with group moves allows a more precise assessment of the importance of the learning

component of the algorithm (repeating good choices and avoiding bad ones, not present in cluster MC algorithms) with respect to the presence of group moves. After optimising the parameter β' of the cluster MC algorithm (see Suppl. S.3.3) we verified that (see Appendix A.1), in a simulation of 2,000,000 steps of 50 (150) particles **1** and **21**, the AB algorithm still finds an energy minimum respectively 22% (35%) and 10% (40%) lower than the cluster MC algorithm. It therefore appears that the performance of the AB algorithm cannot be matched by a simple cluster MC algorithm and that the learning component is indeed essential.

5.4 Conclusion

In this chapter we have proposed and tested a multiscale algorithm that performs a parallel search for the most stable structure that a set of identical particles can form. Using a set of rules, the algorithm identifies "on the fly" stable portions of the system (the agents) in order to reduce the number of steps necessary to reach its lowest energy state. The actions that an agent can perform are move, merge, and split with the addition of the disaggregate action (not present in the lattice version of this algorithm). The decision on which action to perform is based on a set of rules that depend on the configurational energy of the system and on its past memory. The evolution of the system towards the most stable minimum is therefore driven by a combination of adaptation (new configurations require new types of move) and learning (past successful choices should be repeated).

This algorithm is essentially inspired by the analogy between the outcome of molecular self-assembling and emergent phenomena that are studied in the field of complexity. As in many nonmolecular complex systems the emergence of a pattern from the interaction between individuals is best modelled with AB simulations, we propose that an analogous modelling strategy can be successful in the modelling of self-assembling chemical systems. We note that analogy is a standard source of inspiration for the most advanced AI algorithms used in physical sciences, including GA [218] (inspired by evolution in biology) and neural networks [219] (inspired by early works in neurology). As with most optimisation algorithms, it is not possible to identify the system evolution over the course of a simulation with a physical sequence of events, although the agents represent intuitively the most stable portions of the system which can be thought of as the best "bricks" of the simulation for each given configuration. It is of course possible to set up a standard MC simulation at the end of an AB simulation to evaluate thermodynamic averages and free energy. An AB simulation can be seen as an alternative to the equilibration stage of a MC simulation noting however that, in the formation of self-assembly,

equilibration is the most important step and, even when methods that fulfil the detailed balance condition are adopted, the equilibration step cannot be used to derive any thermodynamically relevant parameter.

For a set of arbitrarily shaped particles, we have demonstrated that our rule-based model is able to drive the system to a lower configurational energy in a smaller number of steps than a standard MC simulation. The AB algorithm that includes multiscale moves, that is, both agent moves and single particle moves, finds (on average) an energy minimum 50% lower than the one found by a standard MC method run for the same number of steps. The algorithm has a limited number of parameters that could, in principle, be adjusted on a case-by-case basis. To test the robustness of the algorithm, however, we selected the same sets of algorithm parameters for all the particles considered here, and we therefore verified that the structure of the algorithm is transferable to different types of problem (the performance is likely to improve if the algorithm parameters are optimised for each system).

An essential feature of the proposed method is that all the rules are based on the interaction energy between particles and there are no geometry-related conditions. In this form, the algorithm can be readily extended to any system of rigid particles provided that the total energy can be written as a sum of pairwise interactions, a condition fulfilled by the majority of the force-fields used for studies in solution for the evaluation of nonbonded interactions. Having separated completely the structure of the algorithm from the nature of the self-assembling system makes it possible to use our AB method in conjunction with any molecular modelling package. In the next chapter, we will implement an interface between this algorithm and the freely available TINKER molecular modelling package [220] and the natural continuation of this work will involve the study of realistic molecular systems.

6 AGENT BASED MODELLING OF REALISTIC MOLECULES

"In fact, emergence and complexity theory explains how the principles of self-organization and emergence cause complex systems to arise from simple systems without a top-down designer."

(Michael Shermerr)

Here we extend the AB algorithm, developed in the previous chapter, to the study of the self-assembly of a system of experimental interest. We study the 2D self-assembly of a rigid organic molecule (1,4-benzene-dicarboxylic acid or TPA), comparing the AB results with MC and MC simulated annealing, a technique traditionally used to solve the global minimisation problem. The AB algorithm gives a lower energy configuration in the same simulation time then either of the MC simulation techniques. We also show how the AB algorithm can be used as a part of the protocol to calculate the phase diagram with less computational effort than standard techniques.

6.1 The Minimisation Problem

In general, the prediction of the most stable structure a set of molecule can form involves two main aspects: (i) the development of a reliable force-field able to represent the system under examination and (ii) a modelling strategy able to give the lowest energy configuration of the system described with the chosen force-field. In this chapter we focus on the second aspect, assuming that the available force-fields are capable of giving a reasonable description.

The problem of finding the lowest energy configuration of a set of molecules is analogous to that of finding the global minimum on a very rough multidimensional surface [7]. In this context, MC based methods have proven to be successful [197]. In a classical MC simulation [83], given a configuration s, a new configuration s' is proposed and accepted with the Metropolis probability given by Eq. 2.25. In

a Metropolis MC simulation the system temperature is kept constant throughout the simulation. It is well known that systems simulated with MC will converge to the Boltzmann distribution with a number of simulation steps that cannot be determined *a priori* [102], and the simulations can take a very long time, especially in the case of molecular systems. To overcome this problem, many modifications of the Metropolis MC algorithm have been suggested.

The first simplest modification of the MC method for structure prediction is simulated annealing [103], inspired by the analogy with physical annealing, where a system, initially at a high temperature disordered state, is slowly cooled down becoming more and more ordered. This is the traditional MC-based technique used to solve global minimisation problems [83]. Other advanced MC methods include replica exchange, also known as parallel tempering [104], Wang-Landau [221], and DA [107] algorithms.

Alternative techniques of global optimisation in molecular systems are BH [7], and GAs [113]. In BH a new configuration is generated and minimised, then accepted or rejected following the Metropolis criterion. In this way the real potential energy surface is transformed into a set of plateaus (also called basins) where the barriers between local minima are flattened out. Here, as in the annealing, the limitation is that the simulation should be repeated with many different starting configurations [197]. One way to avoid this is to consider several copies of the same system and perform a parallel search for the lowest energy configuration, as done in GAs. In a GA every replica of the same system is an individual, part of a population, and the population evolves following bio-inspired rules that favour low energy individuals. This method has been used for cluster geometry optimisation [113], crystal structure prediction [116], and protein folding [117].

A further route to solve the global minimisation problem is through AB modelling, a technique originally used to model complex systems in fields such as economics [124], and social sciences [122]. AB models (see Sec. 2.4.3) are rule-based models able to describe the global behaviour of a system on the basis of local interactions between individuals, and in this sense can be considered an extension of CA models [222, 212]. AB models have been recently introduced in chemistry [132, 133, 134, 43, 135], under the assumption that chemical systems behave as complex systems and it is therefore possible to simulate their behaviour with rule based models similar to those used for the study of other complex systems. In chemistry, nucleation and growth of a crystal [132, 133], the packing of a set of molecules to find their lowest energy configuration [43, 135], and the self-organisation of proteins [134] have taken advantage of AB techniques. In particular, our recently developed

AB model for the study of molecular self-organisation [135], presented in the previous chapter, allows a set of shapes to reach their lowest energy configuration through the application of a set of rules.

The aims of this chapter are (i) to extend the algorithm of the previous chapter for the study of realistic systems described by an atomistic force-field, in order to find the lowest energy configuration a set of molecules can form with less computational effort than standard MC techniques, and (ii) to illustrate how the AB algorithm can be successfully used to pre-equilibrate a system, generating configurations that can then be fed into MC simulations to get the statistical averages necessary for the characterisation of the system. As an application, we will construct the phase diagram of the 1,4-benzenedicarboxylic acid (or terephthalic acid, TPA), which is known to form ordered two-dimensional structures on flat surfaces both by gas deposition [**?**, **?**] and at the solid-liquid interface [19]. The simple pattern formed by TPA is one of the most studied examples of multiscale self-assembly: the hydrogen bonds (H-bonds) binding the carboxylic groups in para position bring the molecules together to form chains and the Van der Waals interactions bring the chains close to each other side-by-side. Here we will model the aggregation of TPA on a metallic surface in a ultra-high vacuum environment [**?**, **?**, 226], neglecting the effect of the solvent, although the same pattern was also found in self-assembly from solution [19].

The rest of this chapter is organised as follows: in Sec. 6.2 we describe our AB model and its implementation for the study of realistic systems; the AB model will then be compared to MC simulations in Sec. 6.3 where we also incorporate the AB method in a protocol to compute the phase diagram of TPA. Sec. 6.4 concludes the chapter.

6.2 Model

6.2.1 Agent Based Algorithm

Following Ref. [135], we consider a system composed of N rigid particles (the molecules) in a fixed volume with PBC, at pseudotemperature T (loosely related to the thermodynamic temperature of the system as described below). We identify the *agent* with a collection of particles (a cluster). The size n_i of an agent i is the number of particles contained in the agent. At the beginning of the simulation every agent-cluster will contain only one particle, then each agent will evolve following a set of rules. Each rule is composed of a condition and an action: if the condition is satisfied, the associated action is performed. The conditions in-

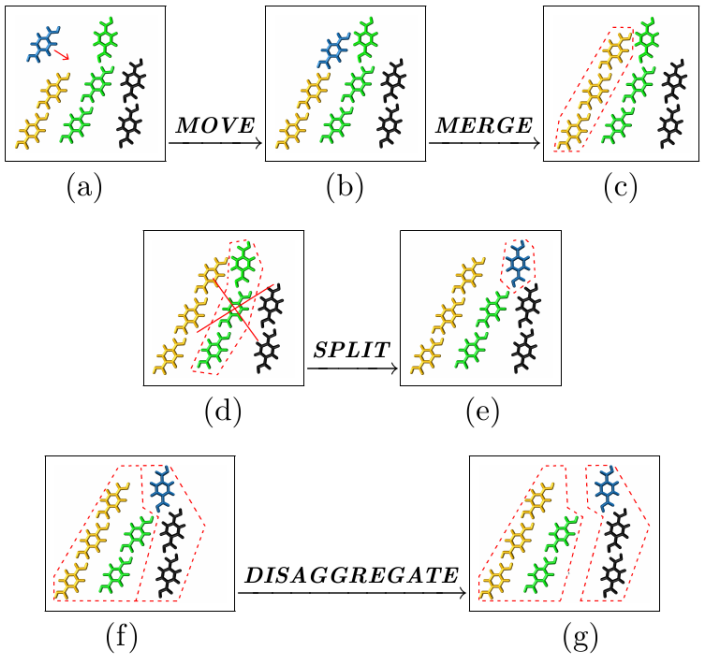


Figure 6.1: The AB algorithm actions, as described in the text. Each agent is represented by a different colour.

volve the comparison between the current configuration of the system and its past history. The descriptors for the current configuration of the system are the agent internal energy $E_i = \frac{1}{2}\sum_{i_k,i_l} V_{i_k i_l}$, where $V_{i_k i_l}$ is the pairwise potential between the two particles i_k and i_l belonging to the i-th agent, and the interaction energy $E_{ij} = \sum_{i_k,j_l} V_{i_k j_l}$ between two agents i and j, where i_k belongs to the agent i, j_l belongs to the agent j, and the summation is taken over all the particles belonging to i and j. The "history" of the system is stored in the vector $E_{\min}[n]$, which keeps track of the internal energy of the most stable agent of each size n formed during the simulation. By using $E_{\min}[n]$ in some of the rules of the algorithm, an element of "learning" is incorporated into the algorithm.

An AB simulation step consists of the successive evaluation of four conditions, each associated with one of the following actions, as schematised in Fig. 6.1: (i) *Move.* A conventional Metropolis MC move, either of the agent or of the single particle, performed at pseudotemperature T (from (a) to (b), Fig. 6.1). If the MC moves were performed only on single particles, or only on clusters of particles that do not change their composition, the system would be rigorously simulated at temperature T. However, because of the other rules, T is not the thermodynamic temperature. (ii) *Merge.* Two agents i and j can merge into a new agent if their

interaction energy E_{ij} is smaller than a reference energy $E_{\rm M}$ (from (b) to (c), in Fig. 6.1). $E_{\rm M} = k_{\rm M} \cdot \{E_{\rm min}[n_i + n_j] - E_{\rm min}[n_i] - E_{\rm min}[n_j]\}$, where the parameter $k_{\rm M} \in [0, 1]$ defines how strict this rule should be (a low value of $k_{\rm M}$ implies a more frequent clustering of "wrong structures"), and the term in the curly brackets is an estimate of the optimal interaction energy between the agents. In addition, the minimum number of agents in the system is set to $N/10$ to allow a comparison amongst agents throughout all the simulations. (iii) *Split.* If the internal energy of an agent i is greater than a certain threshold $E_{\rm S}$, it will split emitting the less bonded particle which will be then defined as a new agent (from (d) to (e), in Fig. 6.1). The new agent will then be brought away by 1000 MC moves with preudotemperature $10 \times T$. $E_{\rm S}$ is set to $k_{\rm S} \cdot E_{\rm min}[n_i]$, where the parameter $k_{\rm S} \in [0, 1]$ determines how often the split rule takes place. (iv) *Disaggregate.* Two agents are part of the same aggregate if their interaction energy is smaller than a certain threshold (i.e., $E_{ij} < E_{\rm D}$, where $E_{\rm D} = k_{\rm D} \cdot \{E_{\rm min}[n_i + n_j] - E_{\rm min}[n_i] - E_{\rm min}[n_j]\}$). If four or more agents are part of the same aggregate, the aggregate is broken by partitioning it into two independent sets of agents [223]. The two sets are then broken apart (from (f) to (g), in Fig. 6.1) by 1000 MC moves with a preudotemperature $10 \times T$. In general, $k_{\rm D} < k_{\rm M}$ and $k_{\rm D} \in [0, 1]$.

The merge rule allows the system to explore space more effectively, by grouping molecules and permitting group moves. The split and disaggregate rules allow the system to escape from kinetic traps. A broader justification of these rules and the method used to set the simulation parameters is given in Ref. [135]. In addition, in the current version of the algorithm, we included the possibility of starting the simulation at higher temperature and gradually decreasing the temperature, in a similar fashion to annealed MC simulations. In the rest of the chapter, AB simulations that use this annealing schedule will be called *annealed AB*. Both agent moves and single particle moves can be performed, and the fraction of single particle moves, which are proven to speed up the convergence [135], is set to the parameter $\chi_{\rm MC}$. The same code is capable of performing MC and annealed MC simulations by switching off all the AB rules. Finally, it is worth noticing that, since the algorithm incorporates an element of "learning" in the rules that depends on its past history, it is substantially different from the MC algorithms with cluster moves presented by us [107] and others [106, 187].

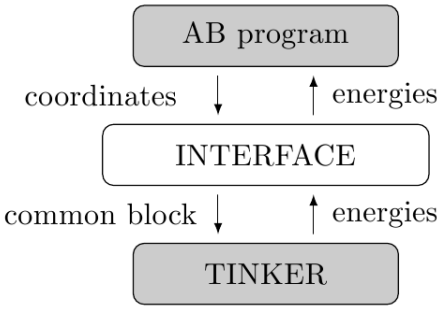


Figure 6.2: To calculate the pair potential $V_{i_k j_l}$ between two molecules, the AB program passes the atomic coordinates of the two molecules i and j to TINKER through an interface. TINKER is then able to calculate the intermolecular energy and passes the information back to the AB program.

6.2.2 Force-Field and Model System

All the AB algorithm conditions are based on energy evaluations which, in turn, are based on pair potential $V_{i_k j_l}$ evaluations. As the vast majority of available force-fields are based on pair potentials, the latter can be calculated with any existing MM code, and the easiest way to link a new program to an existing MM code, without introducing errors in the already established MM model is to build up an *interface*, as shown Fig. 6.2. The interface is needed to keep the AB and MM code structures separated, allowing for an independent development of the two blocks (MM and AB), for example by different research groups. for the MM code, we chose TINKER [220] because it implements several force-fields including the MM3 force-field [74, 224], used in this work for the description of the TPA molecule.

The TPA (Fig. 6.3) molecule, is known to form ordered 2D structures on surfaces [19, 225], as it contains two carboxylic groups in para position that allow the molecules to form H-bonded chains which in turn align side by side held in place by Van der Waals interactions. The distance between the centres of mass of two H-bonded molecules is about 10 Å and two molecules belonging to adjacent chains are separated by 7 Å. It is likely that the surface determines the orientation of the chains in the self-assembly, however, previous studies on surface-free simulations gave structures with the same symmetry and similar lattice parameters to those measured in experiments [226]. Therefore it can be assumed that the function of the metallic surface is only that of keeping the structure planar by acting as a support for the 2D self-assembled structure, allowing us to treat the TPA system as a 2D system. Previous studies of the TPA system with atomistic MD simula-

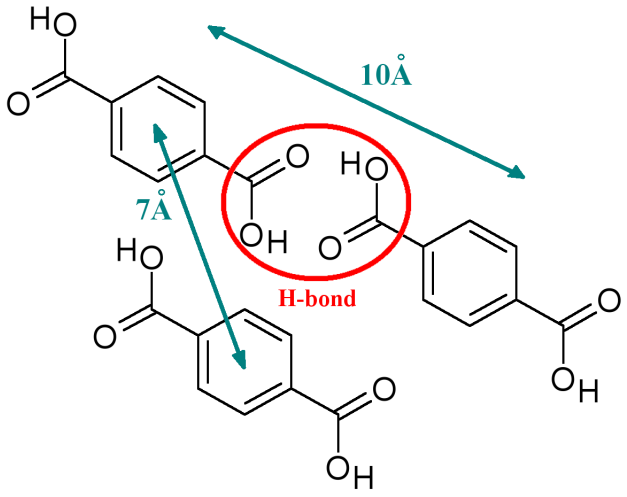


Figure 6.3: Relative position of TPA molecules in the self-assembled structures.

tions [226] and a hexagonal lattice model (see Chap. 4) showed that the molecule can form isolated chains, under dilute conditions, and orderly packed structures, at high density, in agreement with the experimental observations [225]. In this chapter, as a first application of the AB algorithm to a realistic molecule, we will consider a set of rigid and planar TPA molecules in a 2D square box. To keep the system planar, the molecules are allowed to translate on the x, y directions parallel to the plane of the 2D box and to rotate about the z-axis orthogonal to the plane.

The TPA molecules are described with the MM3 force-field (see Apendix A.3). The intermolecular energy contributions (the only one needed) are given by the sum of Van der Waals, H-bonding, and electrostatic interactions. The Van der Waals contributions [74] are given by the Buckingham potential. The H-bonds are calculated over the three atoms involved in the bonding (OH, O), as explained in Ref. [227], and the electrostatic interactions are modelled as a set of bond dipoles [224]. The MM3 functions and parameters are the ones implemented in the TINKER [220] package, without modification. The TINKER default cutoff distance between atoms is set to 9 Å, and, to speed up the algorithm, the energies for the merge and split actions are evaluated only if the distance between the centres of mass of the two molecules is smaller than 15 Å. This is a modification not present in the original algorithm [135], but for realistic systems the estimation of the pair potential is more computationally expensive. The minimum image convention has been used in conjunction with PBC.

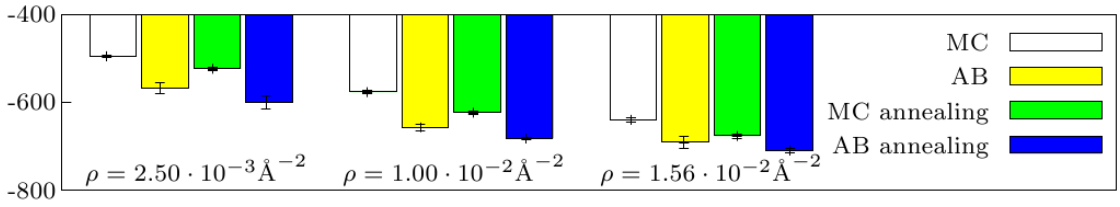


Figure 6.4: Final energy (kcal/mol), averages and errorbars calculated over the last 100 configurations taken every 100 MC steps over 5 runs.

6.3 Results

In this section we will first prove that AB simulations are able to give a lower energy configuration in less simulation time with respect to other MC-based simulation methods. Then, after a few remarks on the AB method, we will apply the AB method to generate a set of starting configurations that will then be equilibrated by MC simulations to get the statistical averages necessary to plot the TPA molecule phase diagram.

6.3.1 Comparison between AB and MC Simulations

We compare here the effectiveness of several algorithms towards the formation of ordered structures. The algorithms considered are AB, MC, annealed MC, and annealed AB. All the simulations are initialised with 100,000 MC steps, then MC simulations have been run for further 1,900,000 steps, and finally a further 900,000 steps of the AB algorithm.

The AB and MC simulations have been run at constant $k_{\mathrm{B}}T = 0.1$ kcal/mol, as this is the temperature that brings the MC simulation to the lowest energy in the given number of steps. In the case of annealed simulations, to identify the temperature range able to give the lowest energy configuration in the given number of steps, we tested MC simulations at different temperatures and we annealed the system starting from $k_{\mathrm{B}}T = 0.2, 0.4, 0.6$ kcal/mol. The lowest configurational energy has been obtained for simulations that started at $k_{\mathrm{B}}T = 0.2$ kcal/mol and ended at $k_{\mathrm{B}}T = 0.1$ kcal/mol, this is the range we will use in this section for the MC annealed and AB annealed simulations, where the schedule of the annealed simulations is set up such that every 20,000 steps for the MC simulation (10,000 for the AB simulation) $k_{\mathrm{B}}T$ is reduced by 0.002 kcal/mol.

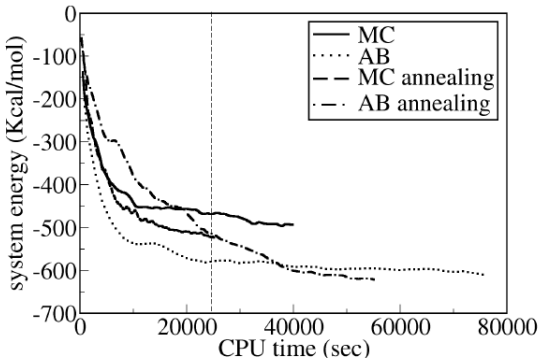
All the simulations were run with a constant number of particles in a 2D square

box of constant size with PBC. We considered systems with 64 molecules and tested several box sizes, equal to 160 Å × 160 Å, 80 Å × 80 Å, 64 Å × 64 Å, giving 2D number densities of $\rho = 2.50 \cdot 10^{-3}, 1.00 \cdot 10^{-2}, 1.56 \cdot 10^{-2}$ Å^{-2}, respectively. Each simulation has been repeated 5 times with different random seeds in order to calculate the average end simulation energies.

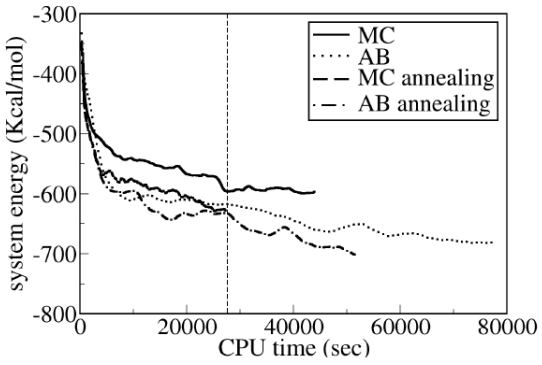
The AB parameter set, introduced in Sec. 6.2.1, has been chosen following the same procedure as in Suppl. S.3.1: we run a set of short simulations (300,000 total simulation steps, after 100,000 initialisation steps) with 50 molecules in a 160 Å × 160 Å square box to identify the parameter set able to give the lowest system energy. We tested χ_{MC} in the range 0.00-0.50, k_{D} in the range 0.10-0.50 and k_{M} in the range 0.70-0.98, with the lowest energy obtained when $\chi_{\mathrm{MC}} = 0.10$, $k_{\mathrm{M}} = 0.70$, $k_{\mathrm{D}} = 0.20$. In addition, we set $k_{\mathrm{S}} = k_{\mathrm{M}}$ without optimising its value. These are the parameters used for all the AB simulations presented in this chapter.

To decide whether the AB simulation is the most suitable simulation scheme to find the lowest energy configuration the system can form, we compared the energy of end simulation configurations and several order parameters. As an AB step is more expensive than a MC step, we run AB simulations for 1,000,000 steps and MC simulations for 2,000,000 steps, as this choice gives similar simulation times for the systems at $\rho = 1.56 \cdot 10^{-2}$ Å^{-2}. Fig. 6.4 summarises the simulation results: for all the considered densities, the AB algorithm is able to reach a lower energy configuration in half the number of steps with respect to MC simulations. Moreover, including the annealing in the AB simulation leads the system to an even lower energy configuration. The AB simulation is also able to reach a lower energy configuration with respect to MC annealing, in the same time (Fig. 6.5), even though at higher density ($1.56 \cdot 10^{-2}$ Å^{-2}) the two algorithms reach a very close energy (Fig. 6.5(c)).

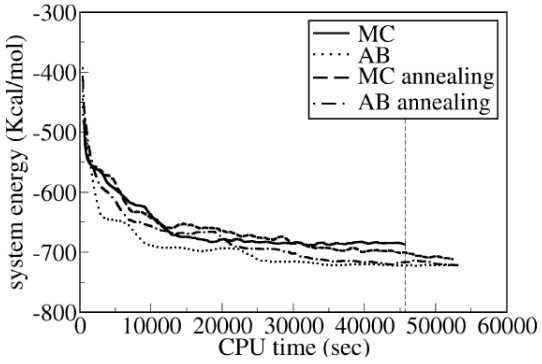
Fig. 6.6 shows snapshots taken at the same simulation time (indicated in Fig. 6.5) of the "best" result for each simulation group, namely configurations of the run that lead to the lowest configurational energy for each simulation type. At low density ($2.50 \cdot 10^{-3}$ Å^{-2}, Fig. 6.6, top row), where we expect to see unidimensional chains (see Sec. 4.3.2), the AB algorithm is able to give the lowest energy configuration with longer chains. Increasing the density, the chains are expected to merge together first to form fibres, then compact structures (as in Fig. 4.12). We instead observe that at intermediate density ($1.00 \cdot 10^{-2}$ Å^{-2}, Fig. 6.6, middle row), only the AB algorithm is able to give 2D structures composed of adjacent parallel chains, while at high density ($1.56 \cdot 10^{-2}$ Å^{-2} Fig. 6.6, bottom row), all the simulation methods give compact structures made up of isoaligned molecules.



(a) $\rho = 2.50 \cdot 10^{-3}$ Å^{-2}



(b) $\rho = 1.0 \cdot 10^{-2}$ Å^{-2}



(c) $\rho = 1.56 \cdot 10^{-2}$ Å^{-2}

Figure 6.5: Comparison between the system energy as a function of the CPU time, for each simulation. The AB simulations have been run for 1,000,000 steps. The MC simulations have been run for 2,000,000 steps. The vertical dashed lines indicate the time at which the snapshots of Fig. 6.6 have been taken.

In the latter case, all the simulations are therefore able to give the target structure, with the AB simulations capable of promoting a higher order (e.g. there are no molecules orthogonal to the average chain direction). To verify quantitatively

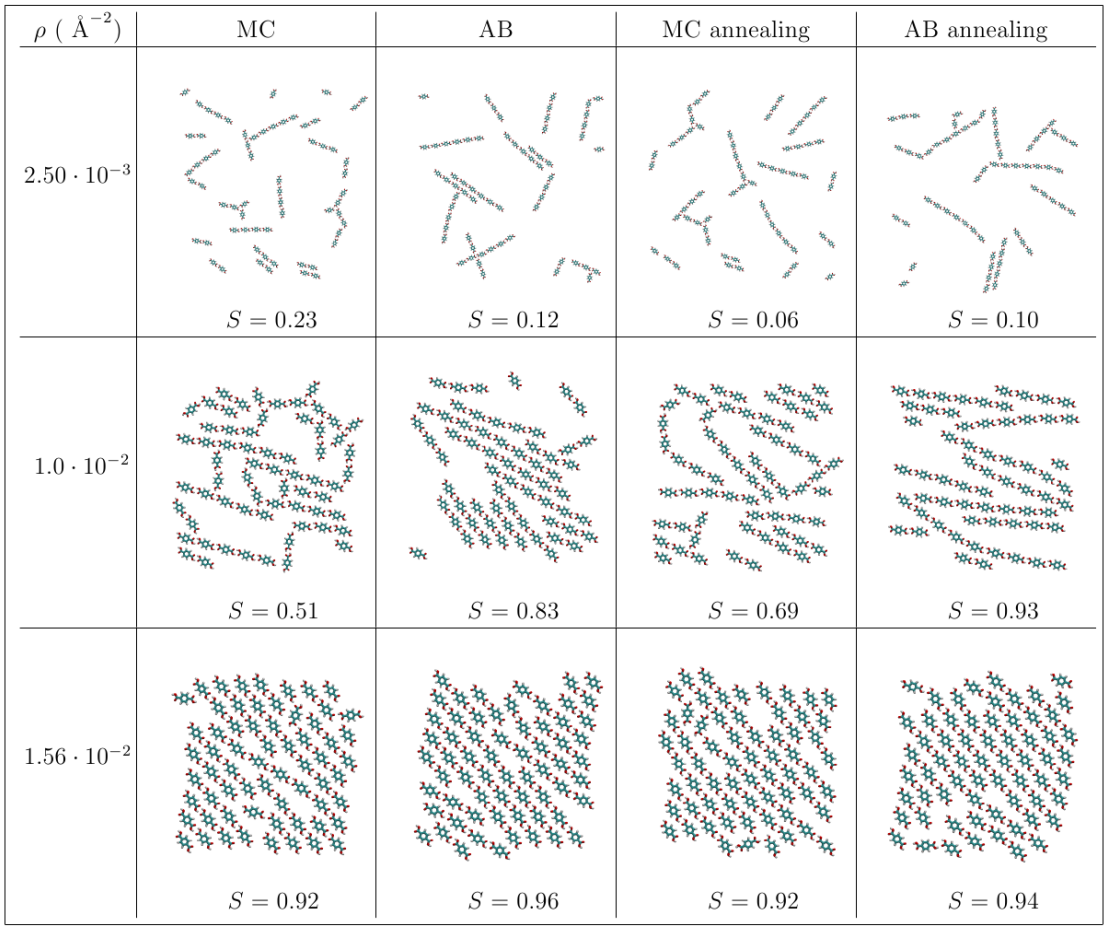


Figure 6.6: TPA end simulation snapshots at different densities, taken at the CPU time indicated by the dashed lines in Fig. 6.5. Snapshots of the simulations that reached the lowest energy configuration for each simulation method are shown. The four columns corresponds to Metropolis MC, AB, annealed MC, and annealed AB simulations.

the latter statement, we assigned to each molecule i an orientation vector e^i, from which it is possible to calculate an order parameter that corresponds to the average orientation of the molecules with respect to a director through the Saupe order tensor Q [228], which in 2D is:

$$Q = 2\left\langle e^i \otimes e^i \right\rangle_i - I \tag{6.1}$$

where $e^i \otimes e^i$ indicates the outer product between the orientation of the i-th molecule with itself, $\langle \cdots \rangle_i$ indicates the average taken over all the molecules, and I is the identity matrix. The eigenvector associated with the largest eigenvalue of Q is

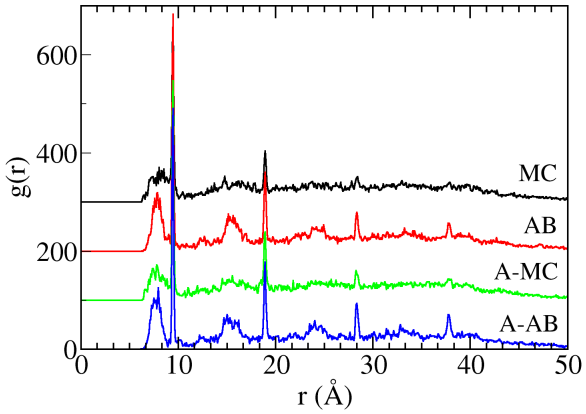


Figure 6.7: Correlation function $g(r)$ for the system with $\rho = 1.00 \cdot 10^{-2}$ Å^{-2}. From top to bottom: MC (black), AB (red), annealed MC (green), annealed AB (blue). For each data set $g(r)$ is off-set by 100. $g(r)$ has been calculated over 100 configurations taken every 100 simulation steps and averaged over 5 runs.

the director of the system, and the eigenvalue associated to it is a measure of the order of the system with respect to the director. The largest eigenvalue of Q is our chosen orientational order parameter S. S has been calculated for the configurations of Fig. 6.6 and indicated under each snapshot. For every system, the AB annealing simulation is able to give a more ordered structure with respect to the MC simulations, as can be seen by comparing the value of S for the given snapshots.

In Fig. 6.7, we report the averaged pair correlation function $g(r)$ computed as:

$$g(r) = \frac{\langle N(r + \Delta r, r)\rangle \cdot A}{\pi[(r + \Delta r)^2 - r^2)] \cdot N} \tag{6.2}$$

where r is the distance between two molecules, $\langle N(r+\Delta r, r)\rangle$ is the average number of molecular pairs at a distance between r and $r + \Delta r$, calculated over several simulation steps and over 5 simulations, N is the total number of particles in the system, and A is the system area. For all the systems the peak at about 10 Å, which corresponds to a H-bonded pair, and the wider peak at 7 Å, which corresponds to two neighboring molecules belonging to adjacent chains (see also Fig. 6.3), are higher for the AB simulations indicating once again that the systems are more ordered.

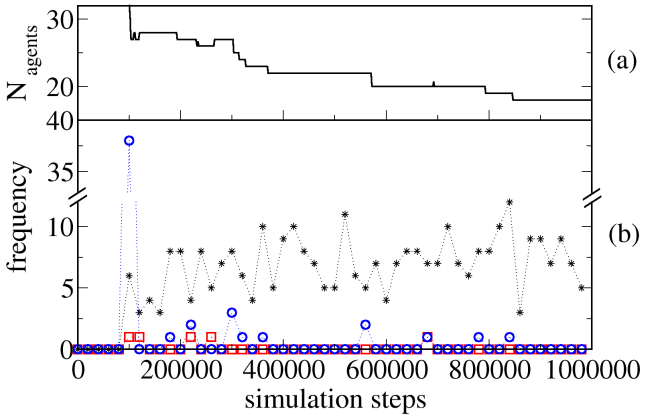
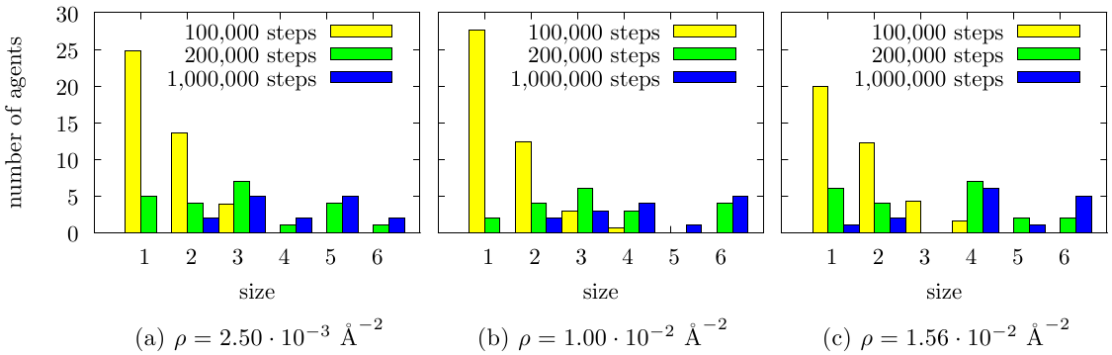


Figure 6.8: Simulation analysis at $\rho = 1.00 \cdot 10^{-2}$ Å^{-2}: (a) number of agents present in the system, (b) frequency of each action calculated every 20,000 steps during the AB simulation: disaggregate (stars), split (squares), merge (circles). The first 100,000 simulation steps represent the initialisation steps, in which only MC moves are performed, the other 900,000 steps are AB steps, in which all actions can be performed.

6.3.2 Agent Based Simulations Analysis

In an AB simulation an agent represents a stable portion of the system under investigation, which optimises itself and grows in time thanks to a set of rules, composed of a set of conditions together with actions to perform if the conditions are satisfied. Fig. 6.8(b) shows how often each action is performed in the AB simulation at $\rho = 1.00 \cdot 10^{-2}$ Å^{-2}. At the very beginning of the simulation, during the 100,000 MC initialisation steps, the algorithm learns which is the most stable dimer it can form and initialises the array $E_{\min}[n]$. After the MC initialisation steps the merge action is the most active (Fig. 6.8(b)) and several clusters form reducing the number of agents present in the system (Fig. 6.8(a)). If initialisation is long enough the split rule does not occur, as we expect to see the formation of only small clusters, mainly dimers. The split rule takes place when larger agents are present which compete to form the most stable cluster and, as the simulation proceeds and the algorithm learns how to form larger stable clusters, the split action no longer occurs. The disaggregate action is the most frequent action throughout all the simulation, it multiplies the attempts of forming larger stable clusters destroying unstable aggregates, leading agents to merge and decreasing the total number of agents in the system (Fig. 6.8(a)).



(a) $\rho = 2.50 \cdot 10^{-3}$ Å^{-2} (b) $\rho = 1.00 \cdot 10^{-2}$ Å^{-2} (c) $\rho = 1.56 \cdot 10^{-2}$ Å^{-2}

Figure 6.9: Number of agent of each size for the best AB simulation after 100,000 (yellow/light gray), 200,000 (green/dark grey), 1,000,000 (blue/black) simulation steps.

It is also possible to analyse the number of agents of each size present in the system at different simulation steps. For every density, the agent size distribution behaves similarly (Fig. 6.9), but denser systems lead to a fast formation of larger agents, as one would expect. At the beginning of the simulation, for all the systems after the 100,000 initialisation steps, dimers and trimers form and a great portion of molecules are still monomeric agents. As the simulation proceeds, monomers slowly disappear merging together or with other agents. After 200,000 simulation steps, the number of monomers is already greatly reduced, and at the end of the simulation monomers are usually not present. In addition, for denser systems, we observe the presence of residual monomers at end-simulation, as they might be encapsulated between agents that have already reached their maximum allowed size. Agent moves are very important for the rearrangement of the agents in order to form larger clusters and it is very active in the first part of the simulation (Fig. 6.10), where we observed the maximum change in the agent size distribution (Fig. 6.9). Moves involving larger agents correspond to many single particle MC moves. The great number of accepted agent moves at the beginning of the simulation contributes to the explanation of why the AB simulation is able to reach faster convergence than MC.

Fig. 6.11a shows how the agents are distributed in the simulation box in an end configuration snapshot. Each agent is formed by two or more aligned molecules that form a chain. The agent-chains are in turn aligned to form the final packed structure. We do not observe agents formed by more than 6 units, due to the restriction employed in the algorithm that the maximum agent size is equal to

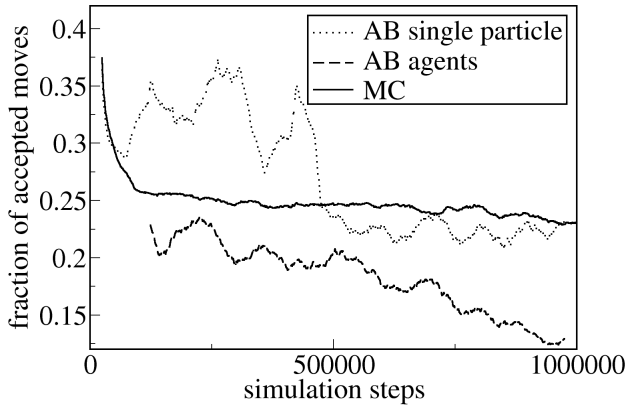
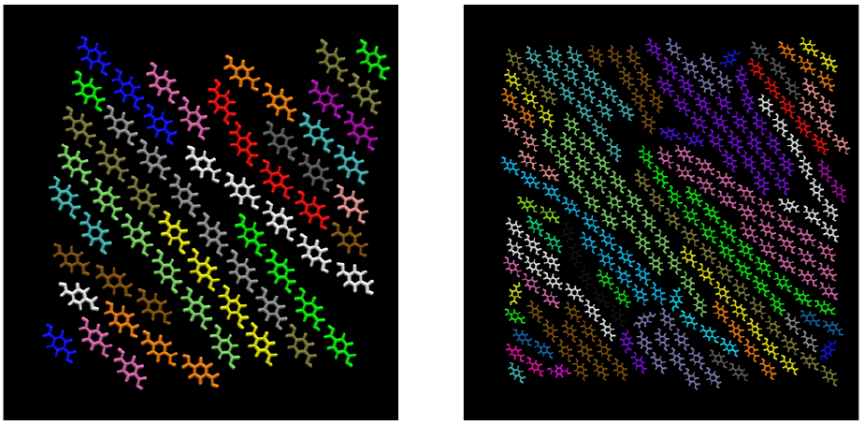


Figure 6.10: Accepted single particle and agent moves for the AB simulation at $\rho = 1.00 \cdot 10^{-2}$ Å^2, and, for comparison, accepted moves for the MC simulation at the same density.

$N/10$, needed to allow the agents to compare themselves with each other and optimise their structure. To see whether larger agents can form by chain alignment, we repeated the simulation for a larger systems composed of 256 molecules in a box of size 128 Å × 128 Å (i.e., at the same high density as in the previous simulations). With this system size, the AB simulations were run for 800,000 steps. We verified that increasing the system size, therefore the maximum agent size, leads the chains to merge together to form larger clusters made up of aligned chains (as in Fig. 6.11b). Moreover, also in this case we obtain a lower configurational energy than a MC simulation run for the same time (11% lower).

6.3.3 TPA Phase Diagram

As shown in the previous sections, the AB algorithm is able to bring a system towards a lower energy configuration with less computational effort than MC simulations. In addition, the AB end configuration structures resemble the true equilibrium structures and are therefore a good starting point for a rigorous simulation. These characteristics make the AB simulation a useful technique to accelerate the equilibration of a system if its thermodynamic properties need to be investigated. However, as the detailed balance is not satisfied in the AB simulation, the MC method should be employed as post-processing to calculate all the statistical averages. As an application, we choose to evaluate the phase diagram of the TPA



(a) (b)

Figure 6.11: End simulation snapshots for different box sizes: (a) 64 Å $\times$ 64 Å and (b) 128 Å $\times$ 128 Å. In both cases the density is $\rho = 1.56 \cdot 10^{-2}$ Å^{-2}. Each colour represents a different agent.

molecule, as this task requires the generation of a very large number of equilibrated configurations.

We will therefore proceed as follows: we use the AB simulation to pre-equilibrate the system, and we then feed the pre-equilibrated configurations into a MC simulation, in order to calculate the necessary thermodynamic quantities. Specifically, we pre-equilibrate 64 molecules systems with 1,000,000 AB steps followed by 100,000 MC steps using the same parameter set as in the previous section, for several combinations of temperature and density. We analyse 600 points in the ranges $T \in [10K, 1000K]$ and $\rho \in [4.44 \cdot 10^{-3}$ Å$^{-2}, 1.56 \cdot 10^{-2}$ Å$^{-2}]$. In terms of CPU time, each simulation has been pre-equilibrated (AB simulation) for 30 hours and run (MC simulation) for 22 hours on a single processor.

First we compare the end configuration energies found by the AB simulation with those explored by subsequent MC steps (Fig. 6.12(a)). At low temperature, the configuration found by AB simulations does not change after subsequent MC steps. At high temperature MC explores a larger configurational space, with the inclusion of higher energy configurations, in the case of high densities, and with the formation of longer (lower energy) chains formed by addition of monomers, in the case of low densities. The difference in energy between MC and AB configurational energies is $\pm 10\%$. However, at high temperatures, the system energy is about -200/-100 kcal/mol (Fig. 6.12(b)) and, given that one H-bonded pair has an energy of about 11 kcal/mol, a 10% difference in energy can be due to the breaking or

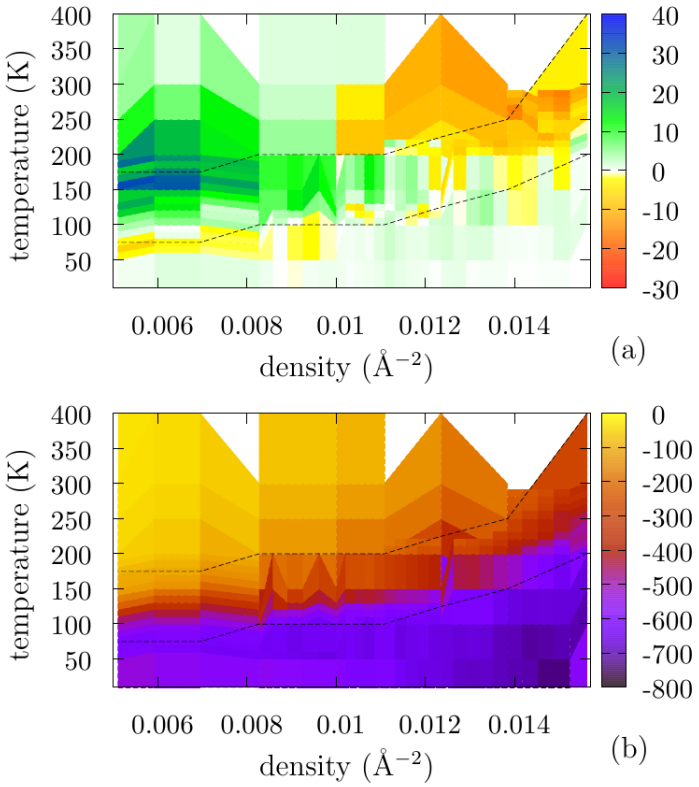


Figure 6.12: (a) difference in energy between the MC and AB simulations, (b) AB end configuration energies. The dashed lines indicate the temperature range in which we observe structural transitions. All the energies are in kcal/mol.

the formation of a double H-bond. This indeterminacy makes it difficult to locate the points of structural transitions with the AB method alone, in particular at low densities. At every density, along the temperature interval in which the energy drops, where structural transitions are supposed to take place, the energies are over/underestimated without regularities.

We observe two phases for this system: a high temperature disordered/fluid phase, which consists of isolated monomers, and a low temperature solid phase which consists of H-bonded chains which can be either isoaligned, forming a packed ordered structure (at high density, as in Fig. 6.6, bottom row), or disordered, forming fibres (at low density, as in Fig. 6.6, top and middle row). To identify the observed phases and the location of the structural transitions between the observed structures we look at the number of double H-bonds per molecule (η_{H}), and the order parameter S (defined in Sec. III.A). Moreover we assign to the H-bonded chains an orientation, from which we calculate the chain order parameter

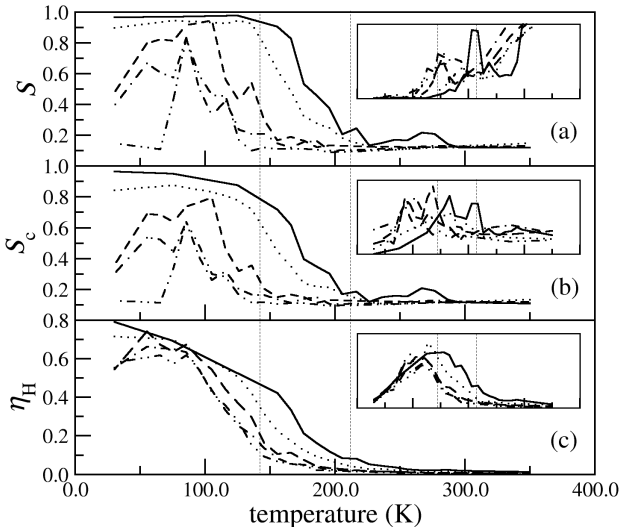


Figure 6.13: Order parameters (a) S, (b) $S_{\rm c}$, and (c) $\eta_{\rm H}$ as a function of the temperature at density d =0.014 Å^{-2} (solid line), 0.012 Å^{-2} (dotted line), 0.010 Å^{-2} (dashed line), 0.008 Å^{-2} (dashed-dotted line), 0.006 Å^{-2} (dashed-dotted-dotted line). In the insets, the variances of the same functions calculated as $\sigma_X = \langle X^2\rangle - \langle X\rangle^2$ in the same temperature range. The vertical dashed line indicates the location of the two structural transitions observed for the system at d =0.014 Å^{-2}.

$S_{\rm c}$, defined analogously to S. For all the order parameters X, we also calculate the variance $\sigma_X = \langle X\rangle^2 - \langle X^2\rangle$, as a peak in these functions may be associated with a structural change.

The order parameters changes with temperature at constant density are plotted in Fig. 6.13 for several densities. At high density (d = 0.0014 $\mathring{A}^{-2}$), we observe a disordered/fluid system at high temperature with no H-bonded pairs present, and an ordered system, with isoaligned chains and with several H-bonded pairs at low temperature. The transition associated with this structural change can be characterised following the variance of the order parameters: at $T \sim$ 200K we observe the onset of order, indicated by a peak in σ_S and $\sigma_{S_{\rm c}}$ and at lower temperature ($T \sim$ 150K) the stabilisation of the H-bonds. At the lowest considered density, the peak of σ_S, $\sigma_{S_{\rm c}}$ and $\sigma_{\eta_{\rm H}}$ are found at the same temperature. And we have a transition between disordered/fluid and fibres.

In order to identify the location of the transition between fibres and packed structure, we plot the same order parameters as a function of the density at constant temperature (Fig. 6.14). Here, at high temperature (e.g. T = 250K) there is a clear transition between the fluid phase and packed structure, as it is possible

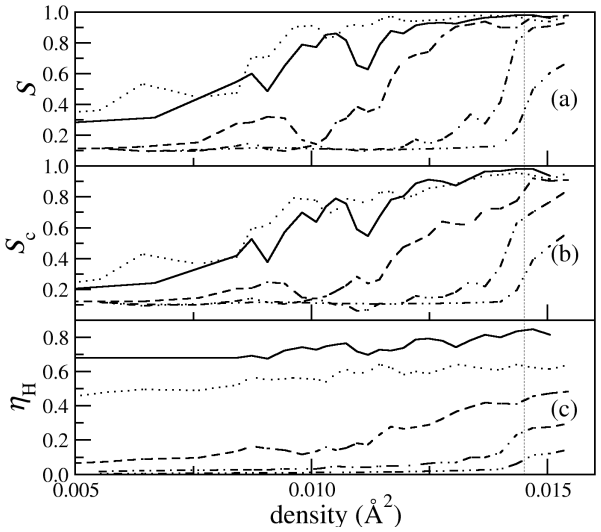


Figure 6.14: Order parameters (a) S, (b) $S_{\rm c}$, and (c) $\eta_{\rm H}$ as a function of the density, at temperature T =50K (solid line), 100K (dotted line), 150K (dashed line), 200K (dashed-dotted line), 250K (dashed-dotted-dotted line). The vertical dashed line indicates the transition from disorder to order at T =250K.

to identify a clear change in all the calculated order parameters. But at lower temperature, where we expect to see a variation in the same parameters due to the structural change between the fibre and packed structures we instead observe a smooth transition.

From the variance of the order parameters at constant density, it is possible to identify the phase boundaries of the system (Fig. 6.15). The T/d phase diagram shows a high temperature fluid phase and a low temperature solid phase in which we observe fibres at low densities and a packed structure at high density. Between the packed and the fluid phases, there is a transient region in which the molecules are aligned but the H-bonds can still form and break. The phase boundaries between the solid phase and the fluid or transient phase are identified by the maxima of $\sigma_{\eta_{\rm H}}$ and $\sigma_{S_{\rm c}}$. The boundaries between the fluid phase and the transient region are identified by the maximum of σ_S. No coexistence region is observed. An absence of a coexistence region has been previously observed in lattice models (as in Chap. 4) and in patchy particle systems [190]. In the latter case, Bianchi and co-workers [190] showed that the absence of a coexistence region is due to the limited number of strong binding sites (or valence) of the self-assembling particles. In particular, as the average valence of the system tends towards two, the gas-liquid coexistence region shrinks and disappears (see Appendix A.4 and Appendix A.5).

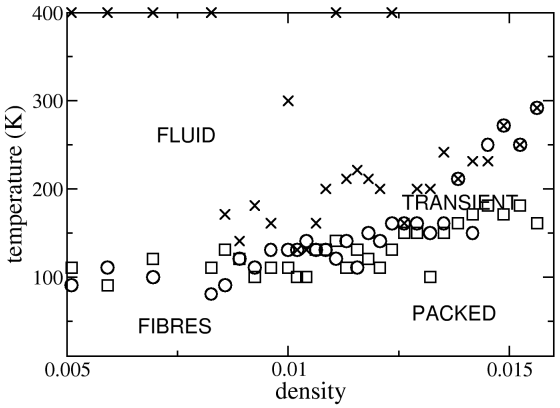


Figure 6.15: TPA phase diagram. The points are the simulation points at which the maximum of the variance of the calculated order parameters have been observed: σ_S (crosses), σ_{S_c} (circles), σ_{η_H} (squares).

6.4 Conclusion

In this chapter, we extended the AB algorithm developed in the previous chapter, which searches the lowest energy configuration a set of molecules can form, to the study of realistic systems described by an atomistic force-field. To this aim, we linked the AB algorithm to TINKER [220], a freely available MM code, and compared the AB algorithm with the Metropolis MC algorithm by running a set of simulations on systems composed of TPA molecules. We proved that the AB algorithm is able to reach lower energy minima with less computational effort, with respect to the Metropolis MC algorithm, also when studying realistic systems. In addition, we incorporated a simulated annealing schedule in the AB algorithm which gives a lower energy configuration with respect to the AB algorithm alone, and both turn out to be more efficient than MC annealing in reaching lower energy configurations.

We then illustrated how the AB algorithm can be successfully used to pre-equilibrate a system, to generate configurations that can then be fed into MC simulations for the proper evaluation of the thermal averages. As an application of the method, we derived the phase diagram of the TPA molecule by approximating the system as a set of rigid molecules constrained into a plane. The obtained TPA phase diagram is consistent with previous observations on systems composed of particles with two strong binding sites [190, 45, 229].

Finally, albeit there is no solvent in the system under investigation, we expect our method to be useful also in the presence of solvent as long as the relative energy of the component molecules can be captured by an effective solvation model [**?**], like the ones commonly implemented in molecular mechanics packages.

In summary, we have shown that the AB methodology can play a role in the molecular modelling of realistic systems as it can be combined with existing force-fields and with more conventional simulation methods such as Metropolis MC and simulated annealing. As for any other tool in the hands of the molecular modeller its usefulness depends partly on the specific problem and partly on the optimisation of the simulation parameters. Through the exploration of different experimental systems and, possibly, alternative sets of rules it will be possible to define more clearly the type of problems for which an AB simulation is the best choice.

7 CONCLUSIONS AND FUTURE WORK

Ad praesens ova cras pullis sunt meliora.

(Latin Proverb)

In this thesis we studied self-organising systems under the assumption that they are complex systems, in an effort to merge together the techniques traditionally used in chemistry, with those typical of complexity science. We assumed that the process of self-organisation could be seen as an emergent phenomena, which is a consequence of the cooperative effects acting among the component of the system. In this framework, in order to find the minimum number of relations able to give the observed patterns, we explored two aspects: (i) the use of minimal models to represent the system under investigation and (ii) the possibility of replacing its real dynamics with a set of rules.

The use of minimal models to represent particle systems, in order to capture the general features of a large class of systems sharing the same characteristics, has a long tradition in physical sciences [48, 44, 46, 70]. In this thesis work, the approach proved once more useful as it allows us to treat large systems and identify the major driving forces responsible for a certain behaviour. This allows us to get a general idea of the expected structures, allowing us to draw phase diagrams which can be used by experimentalists as a guide. We believe that this is a good approach, as force-field models might fail to capture the system-specific differences [226], and experiments still need be performed. However, self-assembling systems are often affected by template effects of the substrate, or by the solvent, this means that it is often necessary to go beyond simplified models in order to get a system-specific knowledge or to adapt the model to the experimental system.

Detailed models can be computational expensive, due to the high cost of the system energy evaluations. To limit this problem, in this thesis we proposed to replace the real dynamics of a self-assembled system with a set of rules. To this aim, we developed an Agent-Based (AB) algorithm for the study of molecular

self-organisation. AB models have been extensively used for the study of complex systems in fields like economics [204] or sociology [205] and were first implemented in chemistry for the study of molecular self-organisation by Troisi *et al.* [43]. The algorithm of Ref. [43], implemented for lattice systems, is a multiscale algorithm that performs a parallel search for the most stable structure that a set of identical particles can form. Using a set of rules, the algorithm clusters the system "on the fly" therefore reducing the number of steps required to reach its lowest energy state. In this thesis we extended the model to the study of realistic systems. To reach this target we first implemented and validated the AB model for off-lattice systems, reformulating its rules and introducing a new rule, namely the disaggregate rule. In addition, we incorporated a simulated annealing schedule in the AB algorithm which gives a lower energy configuration with respect to the AB algorithm alone, and both turn out to be faster than Monte Carlo annealing to reach convergence. We have therefore proven that it is possible to combine the AB scheme with the available molecular modelling algorithms in order to further push down the configurational energy of the system under investigation. Finally, we linked the algorithm to an already available molecular mechanics program, in order to calculate the interaction energies between atomistic molecules. As an application of the method, we derived the phase diagram for a set of rigid molecules constrained in a plane. This is the first example of a rule based model for the study of molecular systems. With the current implementation, the AB model is now an alternative choice to the already available programs.

The power of AB modelling resides in the fact that the rules can be reformulated *ad hoc* in order to achieve the target, which may not necessarily be self-organisation. We believe this is a feasible approach, as in applications it is often unnecessary to simulate the real dynamic of a system, but only its final configuration. The hope is that work will lead to a new area of research. The current AB model can be further improved/modified in several ways: detailed balance could be included, however this will reduce the computational efficiency of the overall algorithm, a minimisation step could be included and, given the possibility of linking the AB algorithm to multiple codes, molecular dynamics (MD) steps could be included as in stochastic MD. Another opportunity includes the addition of multiple resolution schemes. In fact, as AB programs include elements of artificial intelligence, it would be feasible to develop a method to let the algorithm choose the level of approximation with which a portion of the system should be described. Further, we do not exclude the potential extension of the concept of rule-based modelling to reactive systems, or systems composed of different building blocks.

APPENDIX

A.1 Data Augmentation

The result of the AB simulation has been compared with the result given by a DA algorithm, previously implemented in our group [107]. The DA algorithm is an algorithm capable of performing "group moves" respecting the detailed balance. At every DA step, a particle is chosen, and a cluster is defined through a *breadth-first search* algorithm, a graph theory technique able to identify the connected components of a network. The particle network is defined by the probability P_{ij} of two particles to be connected:

$$P_{ij} = 1 - \exp(\beta' \min\{0, E_{ij}\}) \qquad \text{with} \qquad \beta' = k_{\mathrm{B}} T' \tag{A.1}$$

where E_{ij} is the interaction energy between two particles, k_{B} is the Boltzmann constant and T' is a tunable parameter. Once the cluster is defined, it is moved by a random translation or rotation and the move is accepted according to a modified Metropolis acceptance probability:

$$P_{s \to s'} = \min[1, \exp(-[\beta \Delta E - \beta' \Delta E^{-}])] \tag{A.2}$$

where

$$\Delta E = E_{s'} - E_s, \qquad \Delta E^{-} = E^{-}_{s'} - E^{-}_{s}, \qquad E^{-} = \frac{1}{2} \sum_{ij} \min(0, E_{ij}).$$

The modified acceptance probability (Eq. A.2) results from the detailed balance requirement. For more details see Ref. [107].

Fig. A.1 shows the evolution of the total energy of the system as a function of the simulation step for a MC, AB and DA algorithm. The optimisation of the AB and DA parameters is reported in Suppl. S.3.1 and Suppl. S.3.3 respectively. The AB algorithm is capable of driving the system to a lower energy configuration than a DA algorithm.

A.2 Spectral Bisection

The method we employed to divide an aggregate in 2 independent blocks consists in the use of a technique for the *spectral bisection* of a graph [223] and first of all it is necessary to define a connectivity matrix for the components of our system, namely the agents forming the aggregate. To define a connectivity matrix it is

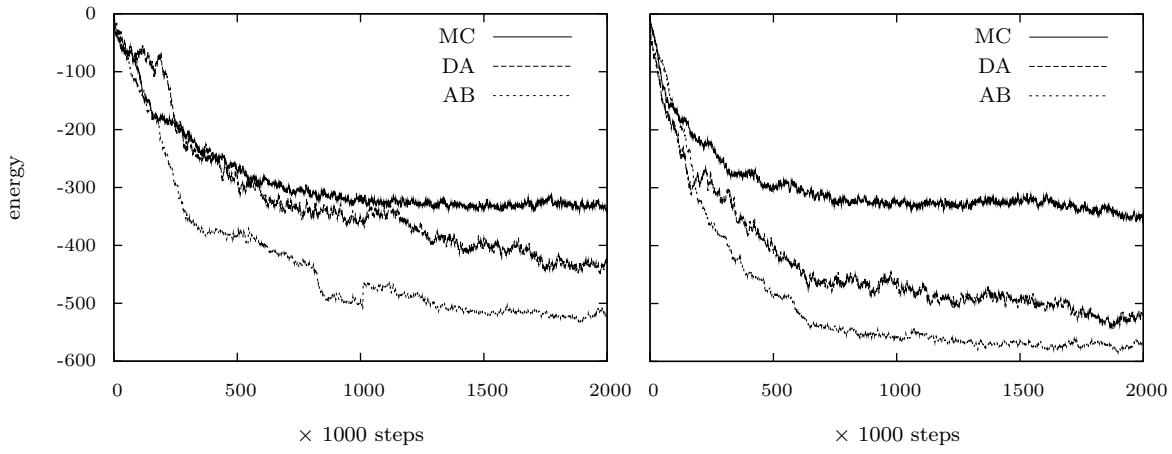


Figure A.1: Performance comparison between MC (solid), DA (dashed), and AB (dotted) simulations: (a) for 150 particles **21** of Fig. 5.3 at $kT = 0.1$, (b) for 150 particles **1** of Fig. 5.3 at $kT = 0.1$.

helpful to define an average interaction energy E_{ref} for the particles forming the aggregate $\mathfrak{A}$:

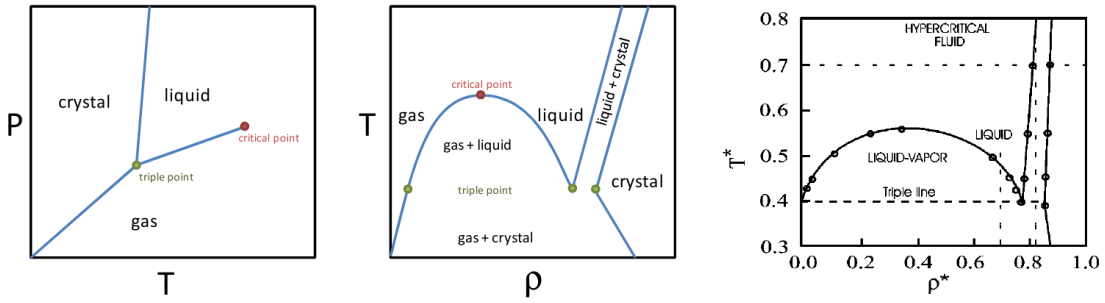
$$E_{\text{ref}} = \frac{\sum_i^{n_{\mathfrak{A}}} \sum_{j<i} E_{ij}}{\sum_i^{n_{\mathfrak{A}}} n_i} \qquad A_i, A_j \in \mathfrak{A} \tag{A.3}$$

where $n_{\mathfrak{A}}$ is the number of agents forming the aggregate $\mathfrak{A}$, E_{ij} is the interaction energy between two agents as defined in Eq. 5.2 and n_i is the size of the i-th agent in the aggregate. This reference value E_{ref} is used to decide whether two agent are connected. Two agents i and j are considered connected if their interaction energy, divided by the sum of the particles composing the agents is lower then E_{ref}. The Laplacian matrix $L(\mathfrak{A})$ is defined as:

$$L_{ij}(\mathfrak{A}) = \begin{cases} \sum_{i'} L_{ii'}(\mathfrak{A}), & \text{if} \quad i = j \\ -1, & \text{if} \quad \frac{E_{ij}}{n_i + n_j} < E_{\text{ref}} \\ 0, & \text{otherwise.} \end{cases} \tag{A.4}$$

Fiedler [230] showed that exists a partition of $\mathfrak{A}$ that tends to give two connected subsystems $\mathfrak{A}^-$ and $\mathfrak{A}^+$ of the original system $\mathfrak{A}$ and the two connected subsystems can be found diagonalising $L(\mathfrak{A})$ and finding its eigenvalues λ_i and eigenvectors v_i. The diagonalisation of the matrix $L(\mathfrak{A})$ has been done with LAPACK [231].

In Ref. [230] it has been proven that, given the series of eigenvalues $\lambda_1 < \lambda_2 <$



(a) (b) (c)

Figure A.2: (a) Schematic p/T phase diagram. (b) Schematic T/ρ phase diagram. (c) Phase diagram of a set of LJ spheres in two-dimensions, adapted from Ref. [232].

$\cdots < \lambda_{n_{\mathfrak{A}}}$, $\lambda_2 \neq 0$ only if the system is connected and the components of its associate eigenvector v_2 define in which partition should be each component of the system:

$$\begin{aligned} &\text{if} \quad v_2(i) < 0 \Rightarrow i \in \mathfrak{A}^- \\ &\text{if} \quad v_2(i) \geq 0 \Rightarrow i \in \mathfrak{A}^+ \end{aligned} \tag{A.5}$$

where $\mathfrak{A}^-$ and $\mathfrak{A}^+$ are the two possible partitions of $\mathfrak{A}$. The theorem states that, using this partition, $\mathfrak{A}^-$ is always connected and, if $v_2(i) = 0, \forall i$, also $\mathfrak{A}^+$ is connected.

A.3 MM3 Force-Field

In the MM3 force-field, the Van der Waals contributions [74] are modelled by the Buckingham potential and depend only on the van der Waals' radii r_v and the potential depth ϵ:

$$E_\mathrm{VdW} = \epsilon\left(1.84 \cdot 10^5 \cdot \exp[-12.00(r/r_\mathrm{v})] - 2.25(r_\mathrm{v}/r)^6\right) \tag{A.6}$$

where r is the interatomic distance. The bond dipoles energies E_dipole are calculated as [224]:

$$E_\mathrm{dipole} = \frac{F\mu\mu'(\cos\chi - 3\cos\alpha\cos\beta)}{R^3 D} \tag{A.7}$$

where $F = 14.39418$ is the conversion factor between ergs/molecule to kcal/mol, μ and μ' are the bond moments of two bonds (in Debyes), χ is the angle between the dipoles, α and β are the angles between the dipoles axes and the lines along

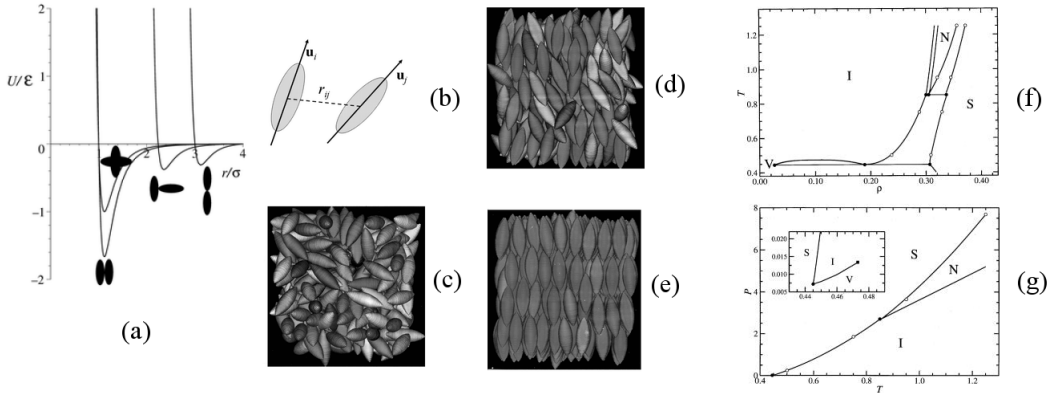


Figure A.3: Potential (a) between Gay-Berne particles (b), particles described with this potential can form isotropic (c), nematic (d) and smectic (d) phases. The observed phases can be collected into a T/ρ (f) and a p/T phase diagram. Figure (a,b,c,d,e) adapted from Ref. [185], (f,g) from Ref. [233].

which the distance between the midpoints of the two bonds (R) is measured, and D is the effective dielectric constant. In addition, the H-bonding potential [227] between an O-H group bonded to a O atom, is given by:

$$E_{\mathrm{HB}} = \frac{\epsilon_{\mathrm{HB}}}{D}\left(1.84 \cdot 10^5 \cdot \exp[-12.00(r/r_{\mathrm{HB}})] - 2.25(r_{\mathrm{HB}}/r)^6 \cdot \frac{l}{l_0}\cos\theta\right) \qquad \text{(A.8)}$$

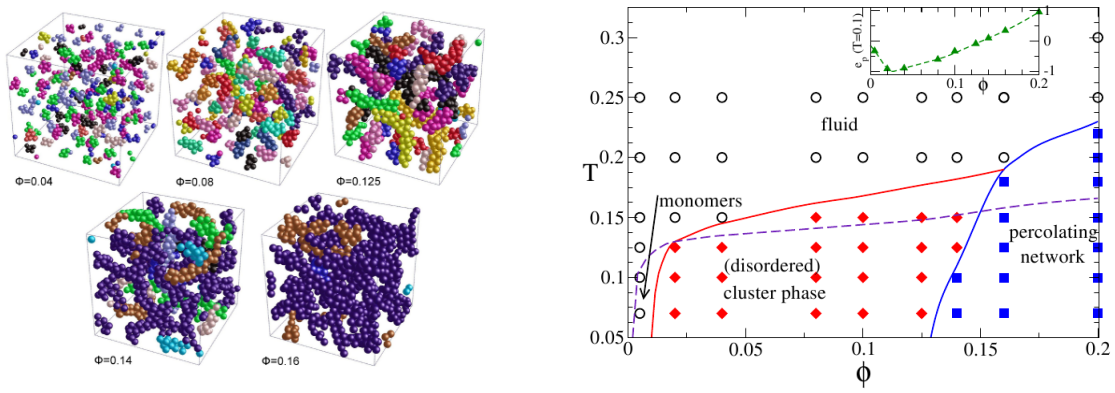
where ϵ_{HB} is the H-bonding parameter in kcal/mol, θ is the angle H-O...O , l is the bond length O-H and l_0 the equilibrium length O-H and r the distance O...H and r_{HB} is the H-bond equilibrium distance.

A.4 Phase Diagrams

The phases expressed by a system are usually schematised into a phase diagram. Two are the most common types of phase diagrams: pressure vs temperature (p/T) or temperature vs density (T/ρ). Simple liquids phase diagrams, like those obtained in the case of LJ spheres [232], shows pure gas/liquid/solid phases separated by coexistence lines in the p/T plane (Fig. A.2(a)) and coexistence regions separating the same pure phases in the T/ρ plane (Fig.A.2(b-c)).

However, as the complexity of a system increases, this simple picture changes its characteristics. To increase the complexity in a system it is possible either to introduce anisotropy in the description of the interactions between building blocks, or to limit the possible number of bonds that particles can form.

In the case of anisotropic potentials, if the possible bonds (or contacts) a



(a) (b)

Figure A.4: Colloidal particles interacting with a LJ plus Yukawa potential, Snapshots as at different packing fractions ϕ (left) and T/ϕ phase diagram (right). Figures taken by Ref. [234]

particle can form are limited only by its shape and geometry, the effect is only that of observing more phases always separated by coexistence regions in the p/T plane. Examples include liquid crystals described with a Gay-Berne potential [185], that show, in addiction to isotropic (Fig. A.3(c)) and crystalline phases, also nematic and smectic phases, which are liquid-crystalline phases. Particles in liquid crystalline phases have long range orientational order, and in the nematic phase (Fig. A.3(d)) their position are not fixed, while in the smectic phase (Fig. A.3(e)) they are organised in ordered layers that can slide upon each other. The effect of the anisotropy on the phase diagrams is therefore only that of including additional phases with their coexistence regions, as shown in Fig. A.3(f-g). Fig. A.3(f) shows the T/ρ phase diagrams of a set of elongated Gay-Berne particles [233] where, with respect to phase diagram for simple fluids (Fig. A.2(b)), an additional phase between the isotropic phase and the solid phase is present together with two coexistence regions. Fig. A.3(g) instead shows the p/T phase diagram for the same Gay-Berne particles: in this diagram multiple triple points are present. This is an example in which the packing determines the possible number of nearest neighbors a particle can have. But if there is a limit in the possible number of bonds that a particle can form, determined by the possible available binding sites (or valence), gel phases can emerge [71].

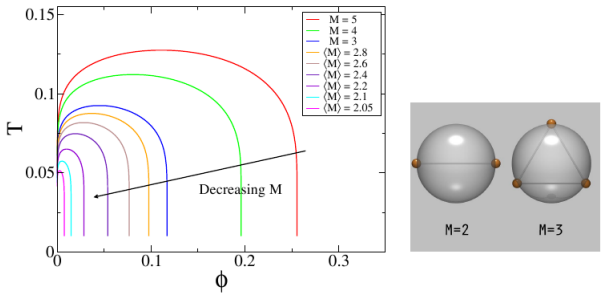


Figure A.5: Shrinking of the coexistence region (left) in the temperature/packing fraction phase diagram for mixture of valence two and three patchy particles (right). Figures taken from Ref. [71].

A.5 Gels Phase Diagrams

Gels are disordered solids, whose components are usually arranged into a network that extends to occupy all the available volume [192, 193]. Gels have been observed in a number of systems composed of particles such as colloids, described for instance with Yukawa potentials [68], patchy particles [190, 71], and models with distinguishable atoms [62]. Most of these studies are however concerned with the low temperature equilibrium structures observed for these systems, and the effect of the potential parameters on the observed phases and only few are concerned with the relations between phases upon changes in the thermodynamic conditions. The difficulty in modelling systems that shows gel phases lies in the fact that their phase diagrams are quite different than those discussed up to now. For example, particles modelled with a LJ plus Yukawa potential [234], show the T/ϕ phase diagram of Fig. A.4(b), where ϕ is the packing fraction. At low temperature, going from low to high packing fraction (Fig. A.4(a)) the particles arrange into monomers, clusters, and finally form a network which extends to occupy all the available space and in which there are no isolated clusters. At high temperature a fluid phase is observed. The most important difference between this phase diagram and the ones in Fig. A.2, is the absence of a coexistence region between the phases. An explanation for this absence comes from the work of Sciortino [190, 71]. Sciortino studied systems composed of patchy particles (Fig. A.5, right) with number of patches M between 2 and 3, and he studied the effect of decreasing the particle valence on the liquid/gas coexistence region simulating mixtures of particles with $M = 2$ and $M = 3$. This work showed that, as the average valence of the system tends towards two, the gas-liquid coexistence region shrinks and disappears. The

absence of a coexistence region allows the formation of dynamical arrested phases such as cluster and network phases. This work therefore proves that networks form thanks to the limited available valence of the building blocks and is not driven by their packing.

Similar phase diagrams, showing a lack of coexistence region, have been be discussed in Chap. 4 and 6, where we will model H-bonded networks formed by molecules that presents two strong binding groups (as H-bonds), extending the study of limited valence systems to a hexagonal lattice model and a realistic system. In particular, in Chap. 6 we will report the first example of computed phase diagram for an atomistic model of molecules with limited valence in two dimension.

SUPPLEMENTAL MATERIAL

S.1 Monte Carlo Simulation

S.1.1 Choice of the Interparticle Potential

In addition to the 6-12 LJ potential and the 14-24 LJ potentials discussed in Chap. 3:

$$E_{\text{LJ 6-12}} = \epsilon_0 \left(\frac{r_{\text{eq}}}{r_{ij}}\right)^{12} - 2 \cdot \left(\frac{r_{\text{eq}}}{r_{ij}}\right)^{6} \tag{S.1}$$

$$E_{\text{LJ 12-24}} = \epsilon_0 \left(\frac{r_{\text{eq}}}{r_{ij}}\right)^{24} - 2 \cdot \left(\frac{r_{\text{eq}}}{r_{ij}}\right)^{12} \tag{S.2}$$

we considered the potentials obtained by summing to those a damped electrostatic contribution (Yukawa potential, with parameters $A = 0.05\epsilon$ 0 , $\xi = 2.00$ taken from [68])

$$E_{\text{Yukawa}} = \frac{A \cdot \exp(-r_{ij}/r_{\text{eq}}/\xi)}{(r_{ij}/r_{\text{eq}}/\xi)} \tag{S.3}$$

Here r_{ij} is the distance between particles i and j (with radius r_i and r_j) and $r_{\text{eq}} = r_i + r_j$. All potentials are given in units of $\epsilon = 100 k_{\text{B}}T$. We also tested a 100-200 LJ potential with a Yukawa contribution (also used sometimes to describe colloidal particles):

$$E_{\text{LJ 100-200}} = \epsilon_0 \left(\frac{r_{\text{eq}}}{r_{ij}}\right)^{200} - 2 \cdot \left(\frac{r_{\text{eq}}}{r_{ij}}\right)^{100} \tag{S.4}$$

A chain of MC simulations has been performed for each potential. Each simulation is linked to the previous one by addition of one particle, going from 188 particles to 242 particles. Each particle is initialised with a random position and a random radius r_j. The distribution of the radii follows a Gaussian distribution with average radius equal to 12.03 nm and standard deviation 1.68 nm, in agreement with experimental determined values. The initial positions are uniformly distributed on a spherical surface with radius $R + r_j$, where $R = 80.79$ nm is the radius of the droplet. In the MC simulation, the maximum displacement on the spherical surface was equal to 0.05rad. The displacement was accepted or rejected following the Metropolis acceptance probability.

Each simulation consists of 50,000,00 MC steps at $k_{\text{B}}T$=0.01, where k_{B} is the Boltzmann constant and T the temperature. The averaged quantities of each simulation have been calculated over the last 25,000,000 steps. 250 configurations have been sampled, one every 100,000 steps. Each chain was repeated 5 times.

The results for the average energies as a function of the number of particles in the system for all the runs are shown in Fig. S.1. From each run an equilibrium configuration has been chosen as the one that minimise the average energy of the system (see Fig. S.1).

We classified the particles of each equilibrium configuration according to their average number of neighbors as described in the text. The results relative to the equilibrium configuration of each potential, averaged over 5 runs, have been compared with the experimental data collected from the experimental micrograph in Fig. S.2. The 6-12 potential, with and without damped electrostatic contribution, underestimates the number of particles with 5 neighbours and overestimates the ones with 6 neighbours. An excellent match is obtained with the 12-24 potential, with and without the Yukawa contribution. The Yukawa contribution to the 6-12 and 12-24 potential has a negligible effect on the neighbours distribution; therefore with those potentials the term can be neglected. No effect is observed comparing the 100-200LJ potential plus Yukawa and the 100-200 potential plus the modified Yukawa potential, which takes into account the charge of particles with equal surface charge density but different size. Both the potentials, being narrow and with a shoulder at distances slightly bigger than the equilibrium distance, does not favour the formation of high coordination numbers and a big excess of particles with 4 neighbours is present with respect to the experimental micrographs. This excess is balanced by the lack of 6 neighboured particles. A reasonable choice of potential to describe the experimental system appears therefore to be the 12-24 LJ potential.

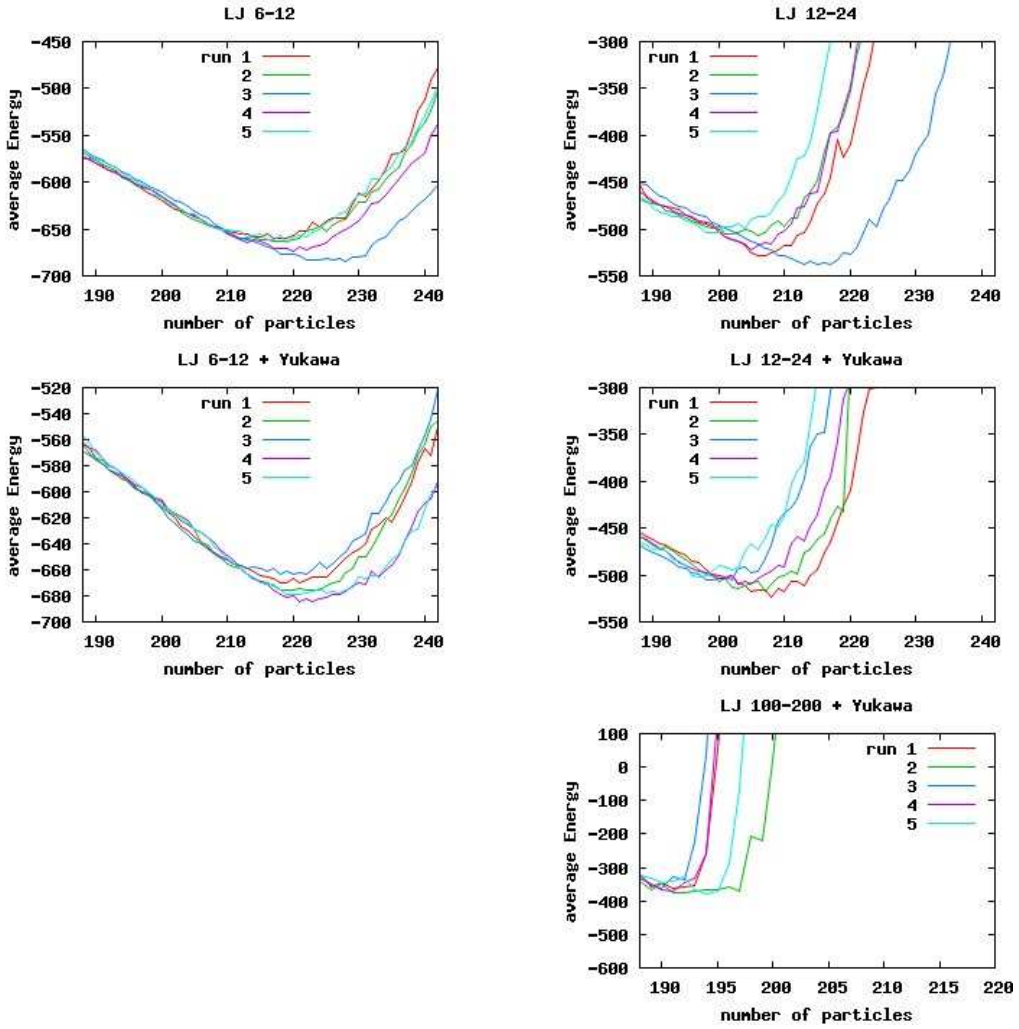


Figure S.1: Average energies as a function of the number of particles on the nanodroplet for a series of different potentials (see text).

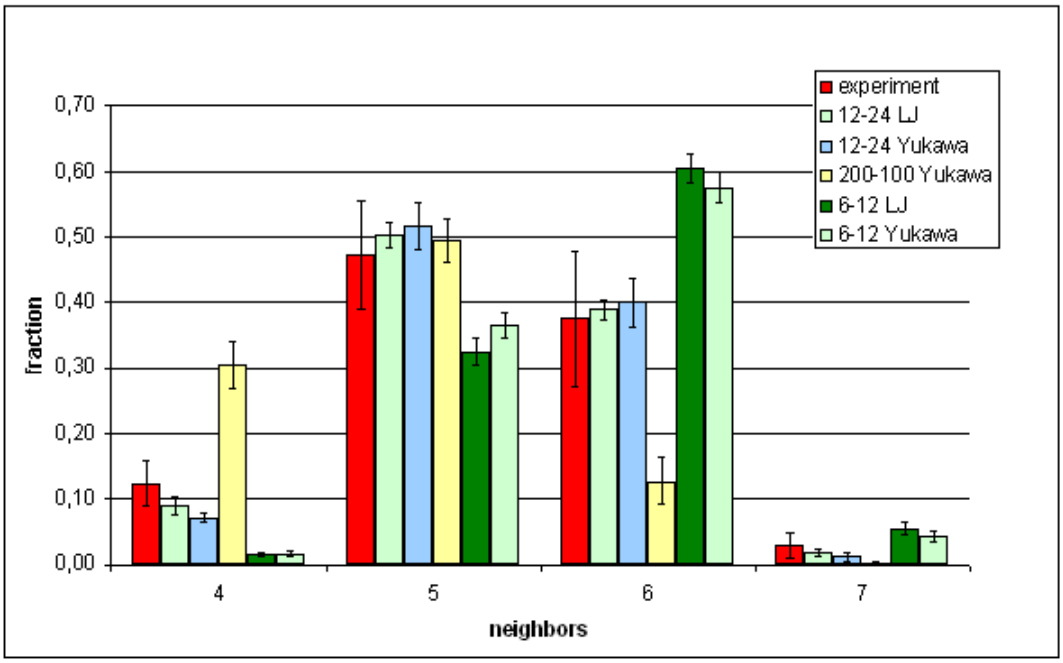
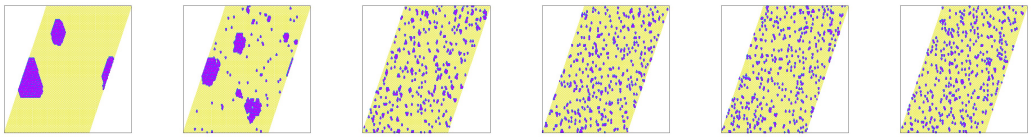


Figure S.2: Nearest neighbour distribution comparison between simulated and experimental data. The best match is obtained for the 12-24 potential both with and without the screen electrostatic contribution.

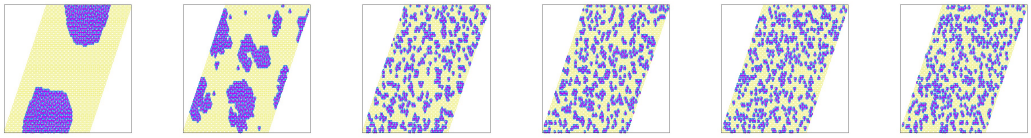
S.2 Hexagonal Lattice Model

S.2.1 Snapshots Tile A



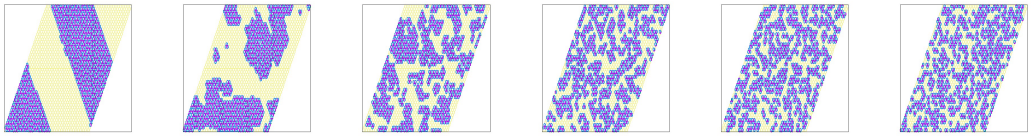
(a) q=1.0 (b) q=1.8 (c) q=2.1 (d) q=2.4 (e) q=2.5 (f) q=3.0

Figure S.3: Tile A, density 0.10



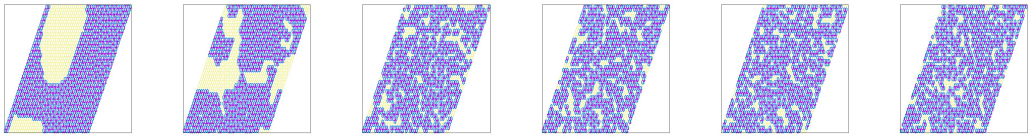
(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

Figure S.4: Tile A, density 0.30



(a) q=1.0 (b) q=1.8 (c) q=2.0 (d) q=2.4 (e) q=3.0 (f) q=4.0

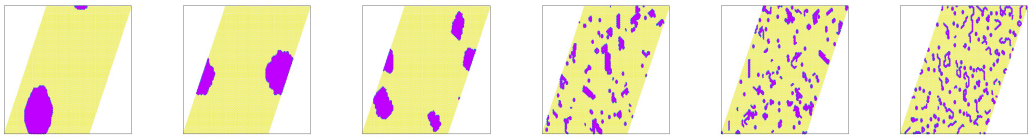
Figure S.5: Tile A, density 0.50



(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

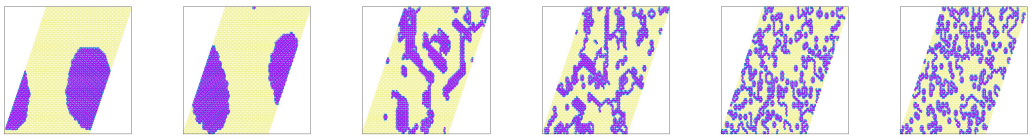
Figure S.6: Tile A, density 0.71

S.2.2 Snapshots Tile B



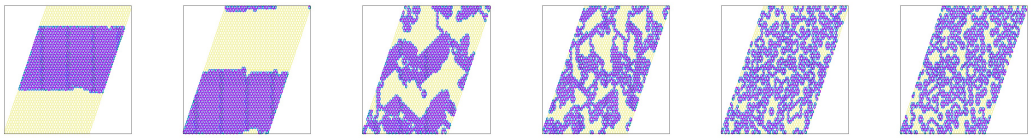
(a) q=1.0 (b) q=1.8 (c) q=2.1 (d) q=2.4 (e) q=2.5 (f) q=3.0

Figure S.7: Tile B, density 0.10



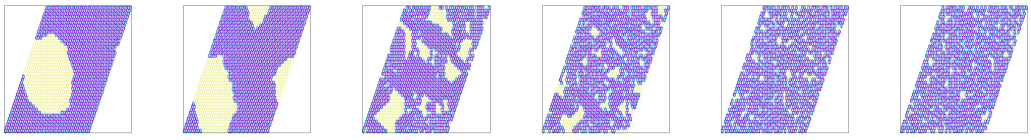
(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

Figure S.8: Tile B, density 0.30



(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

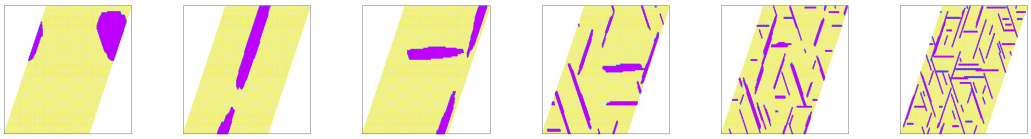
Figure S.9: Tile B, density 0.50



(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

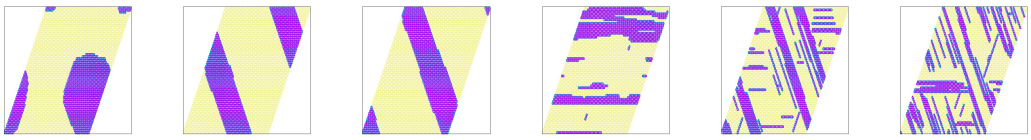
Figure S.10: Tile B, density 0.71

S.2.3 Snapshots Tile C



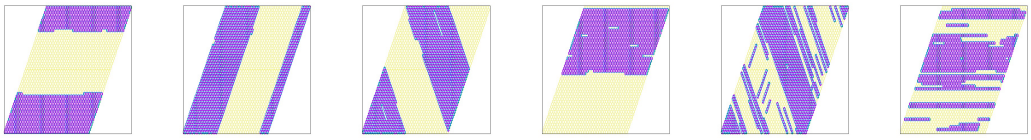
(a) q=1.0 (b) q=1.8 (c) q=2.1 (d) q=2.4 (e) q=2.5 (f) q=3.0

Figure S.11: Tile C, density 0.10



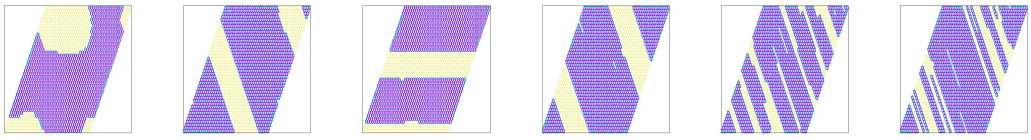
(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

Figure S.12: Tile C, density 0.30



(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

Figure S.13: Tile C, density 0.50



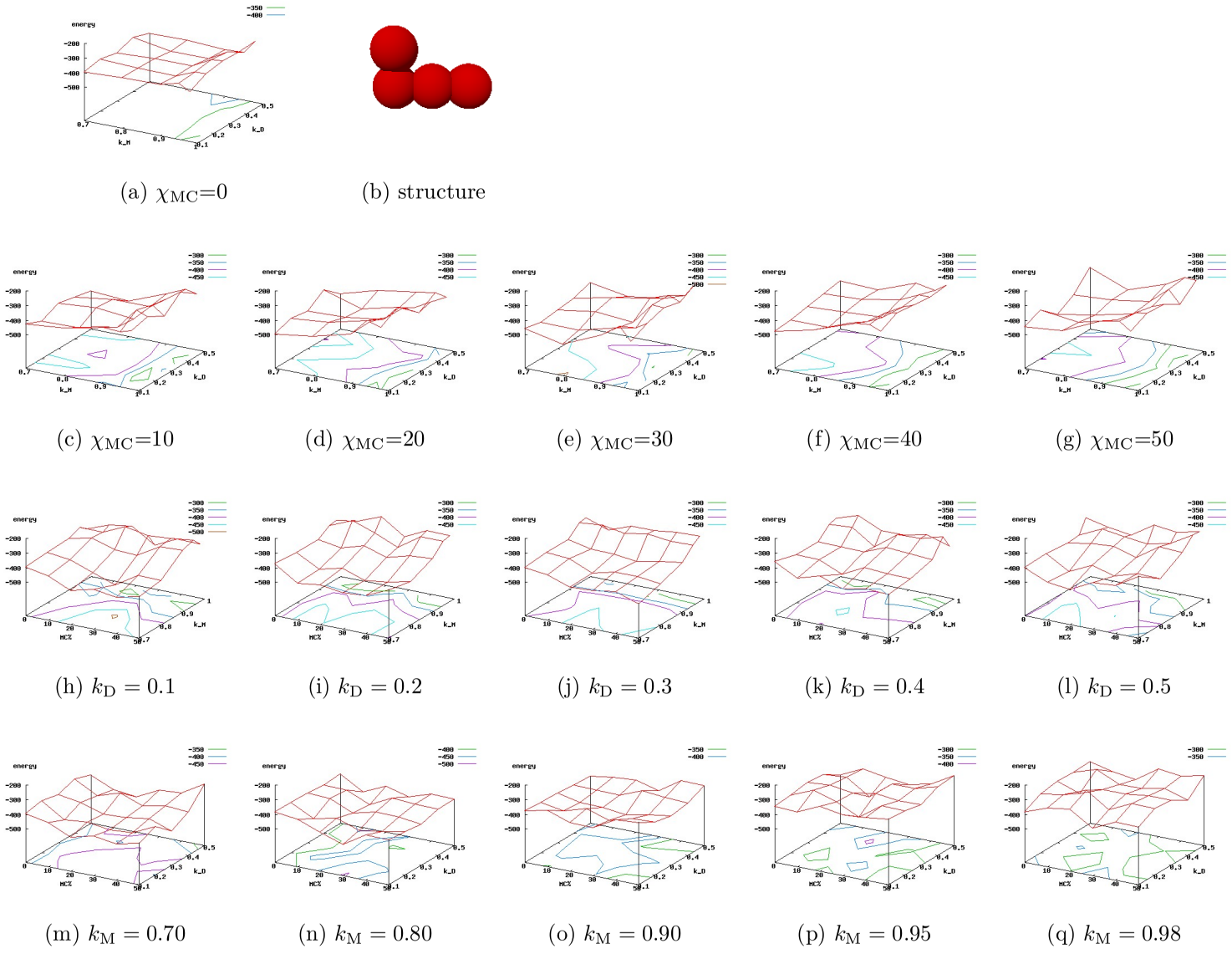
(a) q=1.0 (b) q=1.8 (c) q=2.4 (d) q=2.5 (e) q=3.5 (f) q=4.0

Figure S.14: Tile C, density 0.71

S.3 The Agent Based Model

S.3.1 Agent Based Parameters Study

The parameters studied are χ_{MC} in the range 0.00-0.50, k_M in the range 0.10-0.50 and k_D in the range 0.70-0.98. The best value for χ_{MC} appears to be in the range 0.10-0.50 (see Fig. S.15a and S.15c-S.15g, where χ_{MC} is kept constant and the final energy of the system is evaluated as a function of k_M and k_D, analogously for Fig. S.16a and S.16c-S.16g), this means that it is important to keep the algorithm multiscale including both agent and single particle moves, this assure an internal rearrangement of the agents in order to reach a better structure. The best value for k_D can be chosen as 0.10 (see Fig. S.15h-S.15l, where k_D is kept constant and the final energy of the system is evaluated as a function of χ_{MC} and k_M, analogously for Fig. S.16h-S.16l) and the best value for k_M is 0.80-0.90 (see Fig. S.15m-S.15q, where k_D is kept constant and the final energy of the system is evaluated as a function of χ_{MC} and k_D, analogously for Fig. S.16m-S.16q).



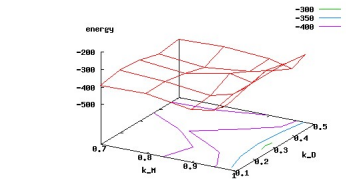
(a) χ_{MC}=0 (b) structure

(c) χ_{MC}=10 (d) χ_{MC}=20 (e) χ_{MC}=30 (f) χ_{MC}=40 (g) χ_{MC}=50

(h) $k_{\mathrm{D}} = 0.1$ (i) $k_{\mathrm{D}} = 0.2$ (j) $k_{\mathrm{D}} = 0.3$ (k) $k_{\mathrm{D}} = 0.4$ (l) $k_{\mathrm{D}} = 0.5$

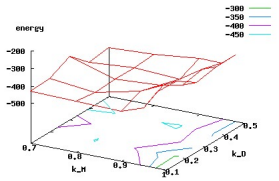
(m) $k_{\mathrm{M}} = 0.70$ (n) $k_{\mathrm{M}} = 0.80$ (o) $k_{\mathrm{M}} = 0.90$ (p) $k_{\mathrm{M}} = 0.95$ (q) $k_{\mathrm{M}} = 0.98$

Figure S.15: Effect of the change in the parameters to the simulation performances for the shape **4** in Fig. 5.3.

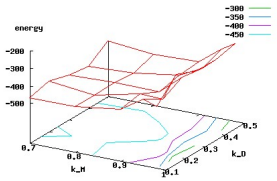


(a) χ_{MC}=0

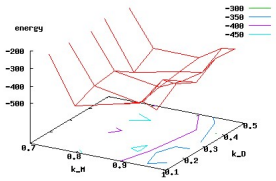
(b) structure



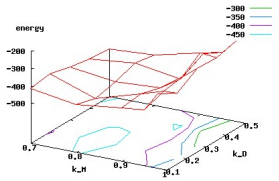
(c) χ_{MC}=10



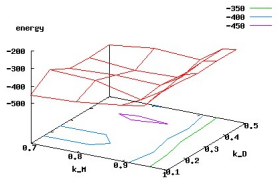
(d) χ_{MC}=20



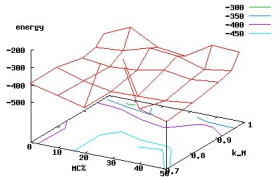
(e) χ_{MC}=30



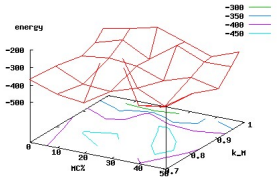
(f) χ_{MC}=40



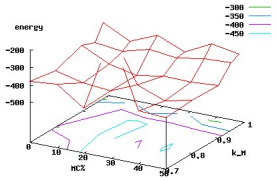
(g) χ_{MC}=50



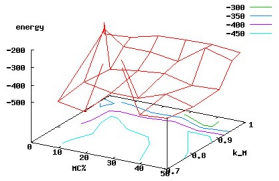
(h) $k_D = 0.1$



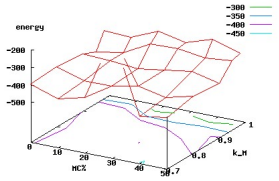
(i) $k_D = 0.2$



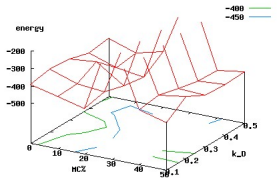
(j) $k_D = 0.3$



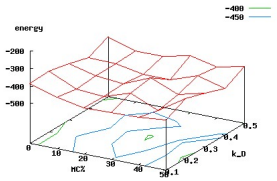
(k) $k_D = 0.4$



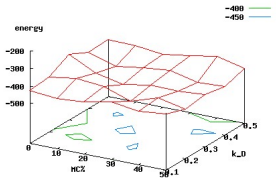
(l) $k_D = 0.5$



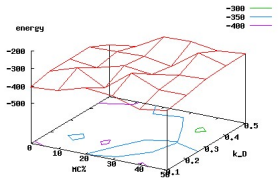
(m) $k_M = 0.70$



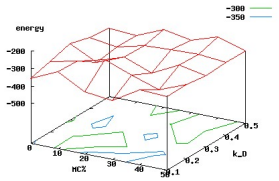
(n) $k_M = 0.80$



(o) $k_M = 0.90$



(p) $k_M = 0.95$



(q) $k_M = 0.98$

Figure S.16: Effect of the change in the parameters to the simulation performances for the shape **26** in Fig. 5.3.

S.3.2 Agent Based Performance with 50 Particles

Two sets of simulations on 50 particles in a box of sie $30 \times 30 \times 30$ with PBC have then be compared: a MC simulation (χ_{MC}=1.00) and AB simulation with only agent moves (χ_{MC}=0.00, $k_{\mathrm{M}} = 0.98$, $k_{\mathrm{D}} = 0.20$) and an AB simulation with single particles moves (χ_{MC}=0.30, $k_{\mathrm{M}} = 0.80$, $k_{\mathrm{D}} = 0.10$).

The performance of the algorithms have been calculated after 2,000,000 simulation steps and compared using the following:

$$\langle R \rangle = \frac{\langle E_{\mathrm{AB}} \rangle - \langle E_{\mathrm{MC}} \rangle}{\langle E_{\mathrm{MC}} \rangle}$$

where $\langle E_{\mathrm{MC}} \rangle$ is the average energy of a MC simulation at equilibrium, and $\langle E_{\mathrm{AB}} \rangle$ is the average energy of the AB simulation. In Tab. S.1 the averages has been calculated over the last 200,000 configurations on 5 runs (one sample every 1000 steps). This table shows that an AB algorithm with only agent moves (χ_{MC}=0.00) reach an energy that is in average the 23% lower then a standard MC simulation, the minimum gain in performance is equal to the 4% for the shape **28** in Fig. 5.3, the maximum gain is equal to 48% for the shape **2** in Fig. 5.3. Including single particle moves has the effect to increase the average performance to the 52% with a minimum of 20% for the shape **28** and a maximum of 78% for the shape **2**.

Another interesting simulation output is the vector $E_{\mathrm{min}}[n]$. The $E_{\mathrm{min}}[n]$ with its standard deviation has been calculated and the data are reported in Tab. S.2 for the runs with only agent moves and in Tab. S.3 for the runs including also single particles moves. Also in the study of this parameter, the inclusion of single particle moves improves the evaluation of $E_{\mathrm{min}}[n]$ and the clusters of each size has been able to reach a lower energy configuration.

shape	$\langle E_{\mathrm{MC}}\rangle \pm \sigma_{E_{\mathrm{MC}}}$ $\chi_{\mathrm{MC}} = 100$	$\langle E_{\mathrm{AB}}\rangle \pm \sigma_{E_{\mathrm{AB}}}$ $\chi_{\mathrm{MC}} = 0$	$\chi_{\mathrm{MC}} = 30$	$\langle R\rangle \pm \sigma_R$ $\chi_{\mathrm{MC}} = 0$	$\chi_{\mathrm{MC}} = 30$
(1)	-304 ± 14	-395± 6	-494 ± 9	0.30 ± 0.06	0.63 ± 0.08
(2)	-280 ± 18	-415± 13	-498 ± 11	0.48 ± 0.10	0.78 ± 0.12
(3)	-321 ± 28	-412± 15	-509 ± 14	0.28 ± 0.12	0.59 ± 0.15
(4)	-372 ± 17	-456± 18	-559 ± 20	0.23 ± 0.07	0.50 ± 0.09
(5)	-375 ± 25	-449± 20	-541 ± 21	0.20 ± 0.10	0.44 ± 0.11
(6)	-384 ± 29	-483± 16	-585 ± 29	0.26 ± 0.10	0.53 ± 0.14
(7)	-379 ± 26	-487± 20	-588 ± 26	0.29 ± 0.10	0.55 ± 0.13
(8)	-371 ± 15	-470± 14	-572 ± 24	0.27 ± 0.07	0.54 ± 0.09
(9)	-363 ± 28	-469± 17	-573 ± 23	0.29 ± 0.11	0.58 ± 0.14
(10)	-378 ± 23	-473± 9	-571 ± 25	0.25 ± 0.08	0.51 ± 0.11
(11)	-404 ± 42	-445± 20	-625 ± 20	0.10 ± 0.13	0.55 ± 0.17
(12)	-408 ± 53	-460± 12	-606 ± 17	0.13 ± 0.15	0.49 ± 0.20
(13)	-417 ± 13	-447± 10	-641 ± 22	0.07 ± 0.04	0.54 ± 0.07
(14)	-455 ± 25	-481± 22	-654 ± 17	0.06 ± 0.08	0.44 ± 0.09
(15)	-416 ± 31	-436± 24	-637 ± 21	0.05 ± 0.10	0.53 ± 0.13
(16)	-351 ± 31	-455± 18	-541 ± 24	0.29 ± 0.13	0.54 ± 0.15
(17)	-371 ± 15	-461± 18	-532 ± 6	0.24 ± 0.07	0.43 ± 0.06
(18)	-346 ± 31	-455± 41	-566 ± 27	0.32 ± 0.17	0.63 ± 0.16
(19)	-366 ± 10	-481± 10	-540 ± 13	0.31 ± 0.04	0.48 ± 0.05
(20)	-362 ± 17	-462± 26	-552 ± 25	0.28 ± 0.09	0.52 ± 0.10
(21)	-365 ± 23	-461± 19	-570 ± 14	0.27 ± 0.10	0.56 ± 0.11
(22)	-387 ± 17	-473± 18	-559 ± 16	0.22 ± 0.07	0.44 ± 0.08
(23)	-381 ± 21	-461± 19	-583 ± 20	0.21 ± 0.08	0.53 ± 0.10
(24)	-353 ± 16	-452± 13	-539 ± 22	0.28 ± 0.07	0.53 ± 0.09
(25)	-366 ± 22	-456± 16	-535 ± 10	0.24 ± 0.09	0.46 ± 0.09
(26)	-399 ± 37	-461± 12	-556 ± 20	0.16 ± 0.11	0.40 ± 0.14
(27)	-371 ± 30	-478± 16	-557 ± 12	0.29 ± 0.11	0.50 ± 0.13
(28)	-455 ± 16	-473± 18	-548 ± 24	0.04 ± 0.05	0.20 ± 0.07
			average	0.23	0.52

Table S.1: Performance comparison, simulation with 50 particles: averaged values calculated over the last 200,000 configurations on 5 runs (one sample every 1000 steps) for each shape of Fig. 5.3. $\langle E_{\mathrm{MC}}\rangle$ and $\langle E_{\mathrm{AB}}\rangle$ are the average value of the system energy for the AB and MC simulations, respectively, and $\sigma_{E_{\mathrm{MC}}}$, $\sigma_{E_{\mathrm{AB}}}$ are their standard deviations. $\langle R\rangle$ has been calculated as in Eq. 5.8 and standard deviation on $\langle R\rangle$ has been calculated as $\sigma_R = \left(\frac{\partial\langle R\rangle}{\partial\langle E_{\mathrm{MC}}\rangle}\right)^2 \sigma^2_{E_{\mathrm{MC}}} + \left(\frac{\partial\langle R\rangle}{\partial\langle E_{\mathrm{AB}}\rangle}\right)^2 \sigma^2_{E_{\mathrm{AB}}}$.

	$\chi_{MC} = 0$, AB%=100			
n	2	3	4	5
shape	$\langle E_{min}[n] \rangle$			
(1)	-7.86±0.06	-17.54±0.62	-26.65±0.96	-34.67±0.40
(2)	-8.23±0.02	-17.28±0.52	-26.01±0.43	-35.11±1.08
(3)	-8.53±0.09	-17.93±0.65	-26.59±0.75	-35.82±0.32
(4)	-9.93±0.03	-21.73±0.00	-33.05±0.00	-42.07±0.67
(5)	-10.24±0.06	-21.42±0.55	-32.24±0.82	-43.83±2.21
(6)	-11.27±0.05	-23.28±0.39	-35.95±2.21	-46.52±1.07
(7)	-10.94±0.07	-22.98±0.73	-33.50±0.64	-44.40±0.91
(8)	-10.47±0.05	-21.71±0.54	-33.22±1.23	-44.59±1.96
(9)	-10.08±0.24	-21.75±0.63	-32.32±1.11	-43.12±1.33
(10)	-10.76±0.02	-23.22±0.87	-33.35±0.81	-43.93±0.88
(11)	-7.86±0.01	-20.86±2.06	-32.59±2.45	-40.06±2.34
(12)	-7.75±0.02	-21.49±0.59	-34.50±2.02	-41.09±3.59
(13)	-8.74±0.01	-22.96±0.28	-34.57±1.75	-42.66±1.80
(14)	-9.08±0.01	-23.10±0.63	-34.08±3.12	-43.85±2.86
(15)	-8.21±0.01	-22.43±0.21	-33.35±2.10	-43.67±4.92
(16)	-10.55±0.06	-21.65±0.44	-32.92±0.78	-43.63±0.47
(17)	-10.55±0.08	-21.25±0.23	-32.25±0.53	-42.80±0.57
(18)	-11.51±0.05	-23.24±0.17	-35.01±0.39	-47.34±1.56
(19)	-11.11±0.02	-22.41±0.19	-34.00±0.59	-45.60±0.45
(20)	-10.57±0.04	-21.30±0.24	-33.09±0.89	-43.52±0.65
(21)	-9.98±0.04	-20.08±0.09	-30.74±0.60	-41.71±1.65
(22)	-10.30±0.03	-20.76±0.17	-31.74±0.92	-43.23±1.27
(23)	-10.61±0.07	-21.29±0.14	-32.15±0.52	-44.09±1.55
(24)	-9.91±0.06	-20.17±0.21	-31.32±1.37	-42.77±1.99
(25)	-10.06±0.07	-21.02±0.77	-30.72±0.56	-42.02±1.05
(26)	-10.78±0.07	-21.63±0.17	-33.62±1.73	-44.56±1.18
(27)	-10.54±0.06	-21.13±0.12	-32.07±0.49	-43.20±1.48
(28)	-10.25±0.05	-20.58±0.21	-32.47±0.87	-43.99±1.46

Table S.2: Average value of $E_{min}[n]$ for the simulations with χ_{MC}=0.00, $k_M = 0.98$, $k_D = 0.20$, $N = 50$; the averages for each n have been calculated over the last 200,000 configurations on 5 runs (one sample every 1000 steps) for each shape of Fig. 5.3.

	χ_{MC} = 30, AB%=70			
n	2	3	4	5
shape	$\langle E_{min}[n] \rangle$			
(1)	-7.97±0.07	-18.18±0.13	-27.79±1.32	-41.69±1.72
(2)	-8.02±0.05	-18.52±0.59	-28.31±1.25	-41.63±0.58
(3)	-8.56±0.11	-18.84±0.38	-30.06±0.99	-42.31±1.13
(4)	-10.39±0.05	-22.87±0.62	-35.60±0.98	-48.63±1.83
(5)	-10.29±0.10	-22.16±0.70	-33.87±1.27	-47.99±0.48
(6)	-11.34±0.05	-23.96±0.38	-36.46±1.31	-49.80±1.36
(7)	-11.00±0.06	-23.89±0.51	-36.25±1.61	-49.70±1.68
(8)	-10.52±0.06	-22.90±0.55	-34.27±1.74	-49.80±2.19
(9)	-10.23±0.06	-22.30±0.65	-34.54±1.56	-48.39±1.39
(10)	-10.79±0.04	-22.95±0.70	-34.97±0.88	-45.92±1.86
(11)	-7.87±0.00	-22.98±0.04	-38.36±0.20	-54.04±0.13
(12)	-7.77±0.00	-22.31±0.12	-37.36±0.29	-52.16±0.46
(13)	-8.76±0.00	-23.73±0.14	-39.96±1.10	-54.68±1.81
(14)	-9.09±0.00	-24.19±0.07	-41.13±0.56	-57.57±0.44
(15)	-8.22±0.01	-23.35±0.18	-38.92±1.13	-54.28±1.44
(16)	-10.65±0.02	-21.56±0.55	-33.88±1.20	-46.68±0.97
(17)	-10.58±0.12	-21.23±0.29	-32.93±1.30	-46.60±2.84
(18)	-11.58±0.06	-23.16±0.11	-36.10±0.91	-48.70±1.43
(19)	-11.12±0.03	-22.25±0.07	-34.79±0.91	-46.22±1.63
(20)	-10.50±0.10	-21.76±0.78	-32.68±0.77	-46.54±2.21
(21)	-10.07±0.05	-20.31±0.12	-32.88±1.48	-49.24±1.41
(22)	-10.30±0.04	-20.68±0.17	-33.34±0.97	-48.72±2.60
(23)	-10.67±0.04	-21.40±0.08	-34.03±0.58	-49.56±1.17
(24)	-9.91±0.03	-22.52±0.43	-33.51±1.68	-47.65±2.37
(25)	-10.14±0.09	-22.10±1.08	-34.73±0.68	-46.12±1.52
(26)	-10.77±0.06	-22.74±0.58	-36.48±1.57	-49.55±0.90
(27)	-10.51±0.06	-21.98±0.71	-34.29±1.65	-46.69±1.42
(28)	-10.25±0.08	-21.82±0.77	-32.62±1.36	-47.41±2.24

Table S.3: Average value of $E_{min}[n]$ for the simulations with χ_{MC}=0.30, $k_M = 0.80$, $k_D = 0.10$, $N = 50$; the averages for each n have been calculated over the last 200,000 configurations on 5 runs (one sample every 1000 steps) for each shape of Fig. 5.3.

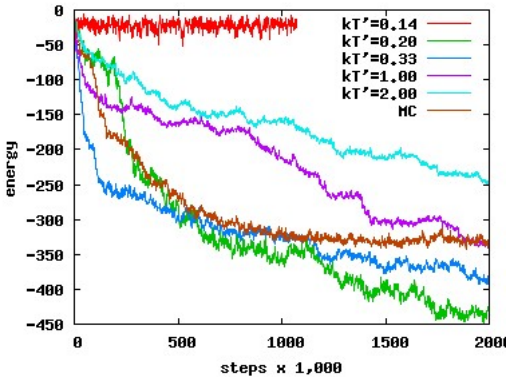


Figure S.17: DA parameters Optimisation. System energy vs. simulation steps, for a DA simulation with selected values of $k_{\mathrm{B}}T' = 0.2$. All the simulations have been run with 50 particles **1** of Fig. 5.3 at $k_{\mathrm{B}}T = 0.1$.

S.3.3 Data Augmentation Parameters Study

We opimise here the DA simulation parameter $k_{\mathrm{B}}T'$, described in Appendix A.1.

We test $k_{\mathrm{B}}T'$ in the range $[0.14, 0.20]$ at $k_{\mathrm{B}}T = 0.1$ in a box of size $30 \times 30 \times 30$ with PBC, in order to identify the which value of $k_{\mathrm{B}}T'$ gives the lowest energy in 2,000,000 simulation steps. The best choice of $k_{\mathrm{B}}T'$ corresponds to $k_{\mathrm{B}}T' = 0.2$ (Fig. S.17).

REFERENCES

[1] Elemans, J., Rowan, A. & Nolte, R. Mastering molecular matter. supramolecular architectures by hierarchical self-assembly. *J. Mater. Chem.* **13**, 2661 (2003).

[2] Angelescu, D. G. & Linse, P. Viruses as supramolecular self-assemblies: modelling of capsid formation and genome packaging. *Soft Matter* **4**, 1981–1990 (2008).

[3] Kato, T., Mizoshita, N. & Kishimoto, K. Functional liquid-crystalline assemblies: Self-organized soft materials. *Angew. Chem.* **45**, 38 (2006).

[4] Barick, K. C. & Bahadur, D. Self-Assembly of Colloidal Nanoscale Particles: Fabrication, Properties and Applications. *J. Nanosci. Nanotechnol.* **10**, 668–689 (2010).

[5] Lehn, J.-M. From supramolecular chemistry towards constitutional dynamic chemistry and adaptive chemistry. *Chem. Soc. Rev.* **36**, 151–160 (2007).

[6] Steed, J. W. & Atwood, J. L. *Supramolecular chemistry* (Wiley, Chichester, 2000).

[7] Wales, D. & Bogdan, E. Potential energy and free energy landscapes. *J. Phys. Chem. B* **110**, 20765 (2006).

[8] Lindsey, J. Self-assembly in synthetic routes to molecular devices. biological principles and chemical perspectives: a review. *New J. Chem.* **15**, 153–180 (1991).

[9] Beer, P., Gale, P. & Smith, D. *Supramolecular Chemistry* (Oxford University Press, Oxford, 1999).

[10] Wales, D. The energy landscape as a unifying theme in molecular science. *philtransa* **363**, 357–375 (2005).

[11] Mohammed, J. S. & Murphy, W. L. Bioinspired design of dynamic materials. *Adv. Mater.* **21**, 2361–2374 (2009).

[12] Ulijn, R. V. & Smith, A. M. Designing peptide based nanomaterials. *Chem. Soc. Rev.* **37**, 664–675 (2008).

[13] Knopp, D., Tang, D. & Niessner, R. Bioanalytical applications of biomolecule-functionalized nanometer-sized doped silica particles. *Anal. Chim. Acta* **647**, 14–30 (2009).

[14] Lu, W. & Lieber, C. M. Nanoelectronics from the bottom up. *Nature Materials* **6**, 841–850 (2007).

[15] Pease, R. & Chou, S. Lithography and other patterning techniques for future electronics. *Proceedings of the IEEE* **96**, 248 (2008).

[16] Blum, A. S. *et al.* Electronic properties of molecular memory circuits on a nanoscale scaffold. *IEEE Trans. Nanobiosci.* **6**, 270–274 (2007).

[17] del Mercato, L. L. *et al.* Charge transport and intrinsic fluorescence in amyloid-like fibrils. *Proc. Nat. Acad. Sci.* **104**, 18019–18024 (2007).

[18] Otero, R. *et al.* Elementary Structural Motifs in a Random Network of Cytosine Adsorbed on a Gold(111) Surface. *Science* **319**, 312–315 (2008).

[19] Lackinger, M., Griessl, S., Markert, T., Jamitzky, F. & Heckl, W. M. Self-assembly of benzene-dicarboxylic acid isomers at the liquid solid interface: Steric aspects of hydrogen bonding. *J. Phys. Chem. B* **108**, 13652–13655 (2004).

[20] Tomba, G. & Vita, A. D. Atomic-level studies of molecular self-assembly on metallic surfaces. *Adv. Mater.* **21**, 1055–1066 (2009).

[21] Feyter, S. D. & Schryver, F. C. D. Two-dimensional supramolecular self-assembly probed by scanning tunneling microscopy. *Chem. Soc. Rev.* **32**, 139–150 (2003).

[22] Barth, J. V., Costantini, G. & Kern, K. Engineering atomic and molecular nanostructures at surfaces. *Nature* **437**, 671–679 (2005).

[23] Stepanow, S., Lin, N., Barth, J. V. & Kern, K. Non-covalent binding of fullerenes and biomolecules at surface-supported metallosupramolecular receptors. *Chem. Comm.* **20**, 2153–2155 (2006).

[24] Theobald, J. A., Oxtoby, N. S., Phillips, M. A., Champness, N. R. & Beton, P. H. Controlling molecular deposition and layer structure with supramolecular surface assemblies. *Nature* **424**, 1029–1031 (2003).

[25] Stepanow, S. *et al.* Steering molecular organization and host-guest interactions using two-dimensional nanoporous coordination systems. *Nature Materials* **3**, 229–233 (2004).

[26] Ebbesen, M. & Jensen, T. Nanomedicine: techniques, potentials, and ethical implications. *J. Biomed. Biotechnol.* **2006**, 51516 (2006).

[27] Kataoka, K., Harada, A. & Nagasaki, Y. Block copolymer micelles for drug delivery: design, characterization and biological significance. *Adv. Drug Deliv. Rev.* **47**, 113–131 (2001).

[28] Duncan, R. The dawning era of polymer therapeutics. *Nat. Rev. Drug Discov.* **2**, 347–360 (2003).

[29] Esfand, R. & Tomalia, D. Poly(amidoamine) (PAMAM) dendrimers: from biomimicry to drug delivery and biomedical applications. *Drug Discov. Today* **6**, 427–436 (2001).

[30] Owens, D. & Peppas, N. Opsonization, biodistribution, and pharmacokinetics of polymeric nanoparticles. *Int. J. Pharm.* **307**, 93–102 (2006).

[31] Douglas, T. & Young, M. Host-guest encapsulation of materials by assembled virus protein cages. *Nature* **393**, 152–155 (1998).

[32] Rosi, N. & Mirkin, C. Nanostructures in biodiagnostics. *Chem. Rev.* **105**, 1547–1562 (2005).

[33] Roy, I. *et al.* Ceramic-based nanoparticles entrapping water-insoluble photosensitizing anticancer drugs: A novel drug-carrier system for photodynamic therapy. *J. Am. Chem. Soc.* **125**, 7860–7865 (2003).

[34] Ow, H. *et al.* Bright and stable core-shell fluorescent silica nanoparticles. *Nano Lett.* **5**, 113–117 (2005).

[35] Binder, W. & Smrzka, O. Self-assembly of fibers and fibrils. *Angew. Chem.* **45**, 7324 (2006).

[36] Nerelius, C., Johansson, J. & Sandegren, A. Amyloid beta-peptide aggregation. What does it result in and how can it be prevented? *Front. Biosci.* **14**, 1716–U3856 (2009).

[37] Deres, K. *et al.* Inhibition of hepatitis B virus replication by drug-induced depletion of nucleocapsids. *Science* **299**, 893–896 (2003).

[38] Day, G. & Motherwell, W. e. a. A third blind test of crystal structure prediction. *Acta Cryst.* **B61**, 511 (2005).

[39] Day, G. M. *et al.* Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test. *Acta Cryst. B* **65**, 107–125 (2009).

[40] Poon, L. & Grebogi, C. Controlling complexity. *Phys. Rev. Lett.* **75**, 4023–4026 (1995).

[41] Nelson, R. *et al.* Structure of the cross-β spine of amyloid-like fibrils. *Nature* **435**, 773 (2005).

[42] Makin, O. S., Atkins, E., Sikorski, P., Johansson, J. & Serpell, L. C. Molecular basis for amyloid fibril formation and stability. *Proc. Nat. Acad. Sci.* **102**, 315–320 (2005).

[43] Troisi, A., Wong, V. & Ratner, M. A. An agent-based approach for modeling molecular self-organization. *Proc. Nat. Acad. Sci.* **102**, 255 (2004).

[44] Baxter, R. J. *Exactly Sovlved Models in Statistical Mechanics* (Academic Press, London, 1982).

[45] Fortuna, S., Cheung, D. L. & Troisi, A. Hexagonal lattice model of the patterns formed by hydrogen-bonded molecules on the surface. *J. Phys. Chem. B* **114**, 18491858 (2010).

[46] Chandler, D. *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, 1987).

[47] Lee, T. D. & Yang, C. N. Statistical theory of equations of state and phase transitions. ii. lattice gas and ising model. *Phys. Rev.* **87**, 410–419 (1952).

[48] Ising, E. Beitrag zur theorie des ferromagnetismus. *Z. Phys. A* **31**, 253–258 (1925).

[49] Onsager, L. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Phys. Rev.* **65**, 117–149 (1944).

[50] Salas, J. The 3-state Potts antiferromagnet on the hexagonal lattice. *J. Phys. A Math. Gen.* **31**, 5969–5980 (1998).

[51] Higo, J., Endo, S. & Nagayama, K. Directional ordering of protein molecules in a two-dimensional hexagonal lattice. an approach with a poker chip model. *Chem. Phys. Lett.* **198**, 300 (1992).

[52] Joknys, A. & Tornau, E. E. Transition order and dynamics of a model with competing exchange and dipolar interactions. *J. Magnetism and Magnetic Materials* **321**, 137–143 (2009).

[53] Barnes, B. C., Siderius, D. W. & Gelb, L. D. Structure, Thermodynamics, and Solubility in Tetromino Fluids. *Langmuir* **25**, 6702–6716 (2009).

[54] Lau, K. & Dill, K. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22**, 3986–3997 (1989).

[55] Krawczyk, J., Owczarek, A. L. & Prellberg, T. Semi-flexible hydrogen-bonded and non-hydrogen bonded lattice polymers. *Physica A* **388**, 104–112 (2009).

[56] Binder, K. & Paul, W. Recent developments in Monte Carlo Simulations of lattice models for polymer systems. *Macromolecules* **41**, 4537–4550 (2008).

[57] Swetnam, A. D. & Allen, M. P. Improved simulations of lattice peptide adsorption. *Phys. Chem. Chem. Phys.* **11**, 2046–2055 (2009).

[58] Paez, A., Tarazona, P., Mateos-Gil, P. & Velez, M. Self-organization of curved living polymers: FtsZ protein filaments. *Soft Matter* **5**, 2625–2637 (2009).

[59] Pasini, P., Chiccoli, C. & Zannoni, C. Liquid crystal lattice models I. Bulk systems. In Pasini, P. & Zannoni, C. (eds.) *Advances in the Computer Simulation of Liquid Crystals*, 99 (Kluwer, Dordecht, 2000).

[60] Jeon, Y. J., Bingzhu, Y., Rhee, T. R., Cheung, D. L. & Jamil, M. Application and new developments in polymer-dispersed liquid crystal simulation studies. *Macrom. Theor. Simul.* **16**, 643–659 (2007).

[61] Tsonchev, S., Niece, K. L., Schatz, G. C., Ratner, M. A. & Stupp, S. I. Phase diagram for assembly of biologically-active peptide amphiphiles. *J. Phys. Chem. B* **112**, 441–447 (2008).

[62] Zhang, Z. & Glotzer, S. Self-assembly of patchy particles. *Nano. Lett.* **4**, 1407–1413 (2004).

[63] Barker, J. & Henderson, D. Theories of Liquids. *Ann. Rev. Phys. Chem.* **23**, 439–484 (1972).

[64] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087 (1952).

[65] Rosenbluth, M. N. & Rosenbluth, A. W. Further results on Monte Carlo equations of state. *J. Chem. Phys.* **22**, 881–884 (1954).

[66] Alder, B. J. & Wainwright, T. E. Phase transition for a hard sphere system. *J. Chem. Phys.* **27**, 1208–1209 (1957).

[67] Wood, W. W. & Jacobson, J. D. Preliminary results from a recalculation of the Monte Carlo equation of state of hard spheres. *J. Chem. Phys.* **27**, 1207–1208 (1957).

[68] Sciortino, F., Mossa, S., Zaccarelli, E. & Tartaglia, P. Equilibrium cluster phases and low-density arrested disordered states: The role of short-range attraction and long-range repulsion. *Phys. Rev. Lett.* **93**, 055701 (2004).

[69] Auer, S. & Frenkel, D. Crystallization of weakly charged colloidal spheres: a numerical study. *J. Phys. Cond. Mat.* **14**, 7667–7680 (2002).

[70] Gay, J. G. & Berne, B. J. Modification of the overlap potential to mimic a linear site–site potential. *J. Chem. Phys.* **74**, 3316–3319 (1981).

[71] Sciortino, F. Gel-forming patchy colloids and network glass formers: thermodynamic and dynamic analogies. *Euro. Phys. J. B* **64**, 505–509 (2008).

[72] Leach, A. L. *Molecular Modelling - Principles and Applications, 2nd Ed.* (Pearson Education Limited, Harlow, 2001).

[73] Allinger, N. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* **99**, 8127–8134 (1977).

[74] Allinger, N. L., Yuh, Y. H. & Lii, J. H. Molecular Mechanics. The MM3 force-field for hydrocarbons. 1. *J. Am. Chem. Soc.* **111**, 8551–8566 (1989).

[75] Weiner, S. J. *et al.* A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784 (1984).

[76] Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

[77] Tozzini, V. Multiscale Modeling of Proteins. *Accounts Chem. Res.* **43**, 220–230 (2010).

[78] Peter, C. & Kremer, K. Multiscale simulation of soft matter systems. *Faraday Discuss.* **144**, 9–24 (2010).

[79] Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. The martini force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824 (2007).

[80] Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **15**, 144–150 (2005).

[81] Vvedensky, D. Multiscale modelling of nanostructures. *J. Phys. Cond. Mat.* **16**, R1537–R1576 (2004).

[82] Newman, M. E. J. & Barkema, G. T. *Monte Carlo Methods in Statistical Physics* (Clarendon Press, Oxford, 1999).

[83] Landau, D. P. & Binder, K. *A guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000).

[84] Perdigão, L. M. *et al.* Functionalized supramolecular nanoporous arrays for surface templating. *Chemistry* **14**, 7600–7607 (2008).

[85] Silly, F. *et al.* Melamine structures on the Au(111) surface. *J. Phys. Chem. C* **112**, 11476–11480 (2008).

[86] Mura, M., Martsinovich, N. & Kantorovich, L. Theoretical study of melamine superstructures and their interaction with the Au(111) surface. *Nanotechnology* **19**, 465704 (2008).

[87] Batchelor, G. K. *An Introduction to Fluid Dynamics* (Cambridge University Press, Cambridge, 1967).

[88] Grant, G. H. & Richards, W. G. *Computational Chemistry* (Oxford University Press, Oxford, 1995).

[89] Adrienko, D., Marcon, V. & Kremer, K. Atomistic simulation of structure and dynamics of columnar phases of hexabenzocoronene derivatives. *J. Chem. Phys.* **125**, 124902 (2006).

[90] Voter, A. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **78**, 3908 (1997).

[91] Maragliano, L. & Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **426**, 168 (2006).

[92] Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Nat. Acad. Sci.* **99**, 12562 (2002).

[93] Barkema, G. T. & Mousseau, N. Event-based relaxation of continuous disordered systems. *Phys. Rev. Lett.* **77**, 4358–4361 (1996).

[94] Malek, R. & Mousseau, N. Dynamics of Lennard-Jones clusters: A characterization of the activation-relaxation technique. *Phys. Rev. E* **62**, 7723–7728 (2000).

[95] Quigley, D. & Rodger, P. M. Metadynamics simulations of ice nucleation and growth. *J. Chem. Phys.* **128**, 154518 (2008).

[96] Quigley, D., Rodger, P. M., Freeman, C. L., Harding, J. H. & Duffy, D. M. Metadynamics simulations of calcite crystallization on self-assembled monolayers. *J. Chem. Phys.* **131**, 094703 (2009).

[97] Laio, A. & Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **71**, 126601 (2008).

[98] Mousseau, N. & Derreumaux, P. Exploring the early steps of amyloid peptide aggregation by computers. *Acc. Chem. Res.* **38**, 885–891 (2005).

[99] Metropolis, N. & Ulam, S. The Monte Carlo Method. *J. Amer. Stat. Ass.* **44**, 335–341 (1949).

[100] Cacciuto, A., Auer, S. & Frenkel, D. Onset of heterogeneous crystal nucleation in colloidal suspensions. *Nature* **428**, 404–406 (2004).

[101] Cacciuto, A., Auer, S. & Frenkel, D. Breakdown of classical nucleation theory near isostructural phase transitions. *Phys. Rev. Lett.* **93**, 166105 (2004).

[102] Narayan, O. & Young, A. P. Convergence of Monte Carlo simulations to equilibrium. *Phys. Rev. E* **64**, 021104 (2001).

[103] Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J. & Caignaert, V. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* **346**, 343–345 (1990).

[104] Swendsen, R. H. & Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **57**, 2607–2609 (1986).

[105] Kofke, D. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.* **117**, 6911–6914 (2002).

[106] Liu, J. & Luijten, E. Rejection-free geometric cluster algorithm for complex fluids. *Phys. Rev. Lett.* **92**, 035504 (2004).

[107] Bhattacharyay, A. & Troisi, A. Self-assembly of sparsely distributed molecules: an efficient cluster algorithm. *Chem. Phys. Lett.* **458**, 210 (2008).

[108] Troisi, A., Wong, V. & Ratner, M. Self-assembly on multiple length scales: A Monte Carlo algorithm with data augmentation. *J. Chem. Phys.* **122**, 024102 (2005).

[109] Guarnieri, F. Theory and Algorithms for mixed Monte Carlo-Stocastic Dynamics Simulations. *J. Mater. Chem.* **18**, 25–35 (1995).

[110] Harmandaris, V. A. & Kremer, K. Dynamics of polystyrene melts through hierarchical multiscale simulations. *Macromolecules* **42**, 791–802 (2009).

[111] Harmandaris, V. A. & Kremer, K. Predicting polymer dynamics at multiple length and time scales. *Soft Matter* **5**, 3920–3926 (2009).

[112] Praprotnik, M., Junghans, C., Site, L. D. & Kremer, K. Simulation approaches to soft matter: Generic statistical properties vs. chemical details. *Comput. Phys. Commun.* **179**, 51 – 60 (2008).

[113] Johnson, R. J. Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. *Dalton Trans.* **22**, 4193–4207 (2003).

[114] Vose, M. D. *The Simple Genetic Algorithm: Foundations and Theory* (The MIT Press, London, 1999).

[115] Jong, K. A. D., Poli, R. & Rowe, J. (eds.) *Foundations of Genetic Algorithms* (Morgan Kaufmann, San Francisco, 2003).

[116] Oganov, A. R. & Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).

[117] Djurdjevic, D. P. & Biggs, M. J. Ab initio protein fold prediction using evolutionary algorithms: Influence of design and control parameters on performance. *J. Comput. Chem.* **27**, 1177–1195 (2006).

[118] Guo, Y., Poulton, G., Valencia, P. & James, G. Designing self-assembly for 2-dimensional building blocks. In *Engineering Self-Organising Systems*, vol. 2977 of *Lect. Notes Comput.Sci.*, 75 (Springer Berlin, Heidelberg, 2004).

[119] Chopard, B. & Droz, M. *Cellular Automata Modeling of Physical Systems* (Cambridge University Press, Cambridge, 1998).

[120] Wolfram, S. Statistical mechanics of cellular automata. *Rev. Mod. Phys.* **55**, 601 (1983).

[121] Packard, N. H. & Wolfram, S. Two-dimensional cellular automata. *J. Stat. Phys.* **38**, 901 (1985).

[122] Bak, P. & Paczuski, M. Complexity, contingency, and criticality. *Proc. Nat. Acad. Sci.* **92**, 6689 (1995).

[123] Main, I. Statistical physics, seismogenesis, and seismic hazard. *Rev. Geophys.* **34**, 433–462 (1996).

[124] Feigenbaum, J. Financial physics. *Rep. Prog. Phys.* **66**, 1611 (2003).

[125] Odagiri, K. & Takatsuka, K. Traveling excitable waves successively generated in a nonlinear proliferation system. *Phys. Rev. E* **79** (2009).

[126] Danielak, R., Perera, A., Moreau, M., Frankowica, M. & Kapral, R. Surface structure and catalytic CO oxidation oscillations. *Physica A* **229**, 428 (1996).

[127] Mukhopadhyay, P., Loeck, M. & Gottstein, G. A cellular operator model for the simulation of static recrystallization. *Acta Mater.* **55**, 551 (2007).

[128] Goetz, R. L. Particle stimulated nucleation during dynamic recrystallization using a cellular automata model. *Scripta Mater.* **52**, 851 (2005).

[129] Yaguma, S., Odagiri, K. & Takatsuka, K. Coupled-cellular-automata study on stochastic and pattern-formation dynamics under spatiotemporal fluctuation of temperature. *Physica D* **197**, 34 (2004).

[130] Wooldridge, M. Intelligent agents: The key concepts. *Lect. Notes Comp. Sci.* **3**, 2322 (2002).

[131] Bonabeau, E., Theraulaz, G. & Deneubourg, J. L. Phase diagram of a model of self-organizing hierarchies. *Physica A* **217**, 373 (1995).

[132] Cartwright, J. H. E., García-Ruiz, J. M., Piro, O., Sainz-Díaz, C. I. & Tuval, I. Chiral symmetry breaking during crystallization: An advection-mediated nonlinear autocatalytic process. *Phys. Rev. Lett.* **93**, 035502 (2004).

[133] Cartwright, J. H. E., Piro, O. & Tuval, I. Ostwald ripening, chiral crystallization, and the common-ancestor effect. *Phys. Rev. Lett.* **98**, 165501 (2007).

[134] Bradford, J. A. & Dill, K. A. Stochastic innovation as a mechanism by which catalysts might self-assemble into chemical reaction networks. *Proc. Nat. Acad. Sci.* **104**, 10098–10103 (2007).

[135] Fortuna, S. & Troisi, A. An artificial intelligence approach for modeling molecular self-assembly: Agent-based simulations of rigid molecules. *J. Phys. Chem. B* **113**, 9877–9885 (2009).

[136] Kondepudi, D. K., Kaufman, R. J. & Singh, N. Chiral symmetry breaking in sodium chlorate crystallization. *Science* **250**, 975–976.

[137] Curl, R. F. & Smalley, R. E. Probing C60. *Science* **242**, 1017–1022 (1988).

[138] Quo, Y., Karasawa, N. & Goddard, W. Prediction of fullerene packing in C60 and C70 crystals. *Nature* **351**, 464–467 (1991).

[139] Wang, L.-L. & Johnson, D. D. Density functional study of structural trends for late-transition-metal 13-atom clusters. *Phys. Rev. B* **75** (2007).

[140] Sara, M. & Sleytr, U. S-layer proteins. *J. Bacteriol.* **182**, 859–868 (2000).

[141] Riebesell, U. *et al.* Reduced calcification of marine plankton in response to increased atmospheric CO_2. *Nature* **407**, 364–367 (2000).

[142] Dinsmore, A. *et al.* Colloidosomes: Selectively permeable capsules composed of colloidal particles. *Science* **298**, 1006–1009 (2002).

[143] Subramaniam, A., Abkarian, M. & Stone, H. Controlled assembly of jammed colloidal shells on fluid droplets. *Nature Materials* **4**, 553–556 (2005).

[144] Giarritta, S. P., Ferrario, M. & Giaquinta, P. Statistical geometry of hard particles on a sphere. *Physica A* **187**, 456 – 474 (1992).

[145] Thomson, J. J. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *Philos. Mag.* **7**, 237–265 (1904).

[146] Tammes, R. On the origin number and arrangement of the places of exits on the surface of pollengrains. *Rec. Trv. Bot. Neerl.* **27**, 1–84 (1930).

[147] Bausch, A. *et al.* Grain boundary scars and spherical crystallography. *Science* **299**, 1716–1718 (2003).

[148] Lipowsky, P., Bowick, M., Meinke, J., Nelson, D. & Bausch, A. Direct visualization of dislocation dynamics in grain-boundary scars. *Nature Materials* **4**, 407–411 (2005).

[149] Bon, S. A. F. & Colver, P. J. Pickering miniemulsion polymerization using Laponite clay as a stabilizer. *Langmuir* **23**, 8316–8322 (2007).

[150] Cauvin, S., Colver, P. & Bon, S. Pickering stabilized miniemulsion polymerization: Preparation of clay armored latexes. *Macromolecules* **38**, 7887–7889 (2005).

[151] Weiss, K. Paint and coatings: A mature industry in transition. *Prog. Polym. Sci.* **22**, 203–245 (1997).

[152] Liu, H. *et al.* One-pot fabrication of magnetic nanocomposite microcapsules. *Mater. Lett.* **63**, 884–886 (2009).

[153] Ramsden, W. Separation of solids in the surface-layers of solutions and 'suspensions' (observations on surface-membranes, bubbles, emulsions, and mechanical coagulation) - preliminary account. *Proc. R. Soc. London, Series A* **72**, 156–164 (1903).

[154] Pickering, S. U. Emulsions. *J. Chem. Soc., Trans.* **91**, 2001–2021 (1907).

[155] Pieranski, P. Two-dimensional interfacial colloidal crystals. *Phys. Rev. Lett.* **45**, 569–572 (1980).

[156] Danov, K., Kralchevsky, P., Ananthapadmanabhan, K. & Lips, A. Particle-interface interaction across a nonpolar medium in relation to the production of particle-stabilized emulsions. *Langmuir* **22**, 106–115 (2006).

[157] von Smoluchowski, M. Drei Vorträge über Diffusion, Brownsche Molekularbewegung und Koagulation von Kolloidteilchen. *Phys. Z.* **17**, 557–571 (1916).

[158] Abkarian, M. *et al.* Dissolution arrest and stability of particle-covered bubbles. *Phys. Rev. Lett.* **99**, 188301 (2007).

[159] Zhang, J., Blaak, R., Trizac, E., Cuesta, J. A. & Frenkel, D. Optimal packing of polydisperse hard-sphere fluids. *J. Chem. Phys.* **110**, 5318–5324 (1999).

[160] Voivret, C., Radjaï, F., Delenne, J.-Y. & El Youssoufi, M. S. Space-filling properties of polydisperse granular media. *Phys. Rev. E* **76**, 021301 (2007).

[161] Viveros-Mendez, P. X., Mendez-Alcaraz, J. M. & Gonzalez-Mozuelos, P. Two-body correlations among particles confined to a spherical surface: Packing effects. *J. Chem. Phys.* **128**, 014701 (2008).

[162] Bowick, M., Cacciuto, A., Nelson, D. R. & Travesset, A. Crystalline order on a sphere and the generalized Thomson problem. *Phys. Rev. Lett.* **89**, 185502 (2002).

[163] Li, C. R., Dong, W. J., Gao, L. & Cao, Z. X. Stressed triangular lattices on microsized spherical surfaces and their defect management. *Appl. Phys. Lett.* **93**, 034108 (2008).

[164] Renka, R. Algorithm 772: STRIPACK: Delaunay triangulation and Voronoi diagram on the surface of a sphere. *ACM T. Math. Software* **23**, 416 (1997).

[165] Fowler, P. W. & Manolopoulos, D. E. *An Atlas of Fullerenes* (Clarendon Press, Oxford, 1995).

[166] Stepanow, S. *et al.* Programming supramolecular assembly and chirality in two-dimensional dicarboxylate networks on a Cu(100) surface. *Nano. Lett.* **5**, 901–904 (2005).

[167] Dou, W. *et al.* Aggregation of organic molecules on silver surface with the balance between molecule-substrate interaction and intermolecular interaction. *Chem. Phys. Lett.* **470**, 126 – 130 (2009).

[168] Heininger, C., Kampschulte, L., Heckl, W. M. & Lackinger, M. Distinct differences in self-assembly of aromatic linear dicarboxylic acids. *Langmuir* **25**, 968–972 (2009).

[169] Zhu, N., Osada, T. & Komeda, T. Supramolecular assembly of biphenyl dicarboxylic acid on Au(111). *Surf. Sci.* **601**, 1789–1794 (2007).

[170] Kwon, K.-Y., Lin, X., Pawin, G., Wong, K. & Bartels, L. Oxadiazole-metal interface: from isolated molecules to π-stacking. *Langmuir* **22**, 857–859 (2006).

[171] Miyashita, N. & Kurth, D. G. Directing supramolecular assemblies on surfaces. *J. Mater. Chem.* **18**, 2636–2649 (2008).

[172] Xiao, W. *et al.* Self-assembly of chiral molecular honeycomb networks on Au(111). *J. Am. Chem. Soc.* **130**, 8910–8912 (2008).

[173] Madueno, R., Raisanen, M. T., Silien, C. & Buck, M. Functionalizing hydrogen-bonded surface networks with self-assembled monolayers. *Nature* **454**, 618 (2008).

[174] Zwaneveld, N. A. A. *et al.* Organized formation of 2d extended covalent organic frameworks at surfaces. *J. Am. Chem. Soc.* **130**, 6678–6679 (2008).

[175] Wang, Y. *et al.* Azo supramolecules on Au(111) with controlled size and shape. *J. Am. Chem. Soc.* **130**, 4218–4219 (2008).

[176] Yokoyama, T., Yokoyama, S., Kamikado, T., Okuno, Y. & Mashiko, S. Selective assembly on a surface of supramolecular aggregates with controlled size and shape. *Nature* **413**, 619–621 (2001).

[177] Clair, S. *et al.* Stm study of terephthalic acid self-assembly on au(111): Hydrogen-bonded sheets on an inhomogeneous substrate. *J. Phys. Chem. B* **108**, 14585–14590 (2004).

[178] Okuno, Y., Yokoyama, T., Yokoyama, S., Kamikado, T. & Mashiko, S. Theoretical study of benzonitrile clusters in the gas phase and their adsorption onto a Au(111) surface. *J. Am. Chem. Soc.* **124**, 7218–7225 (2002).

[179] Rapino, S. & Zerbetto, F. Modeling the stability and the motion of DNA nucleobases on the gold surface. *Langmuir* **21**, 2512–2518 (2005).

[180] Klappenberger, F. *et al.* Does the surface matter? Hydrogen-bonded chain formation of an oxalic amide derivative in a two- and three-dimensional environment. *ChemPhysChem* **9**, 2522–2530 (2008).

[181] Silly, F. *et al.* Deriving molecular bonding from a macromolecular self-assembly using kinetic Monte Carlo simulations. *Phys. Rev. B* **77**, 201408 (2008).

[182] Kampschulte, L. *et al.* Thermodynamical equilibrium of binary supramolecular networks at the liquid - solid interface. *J. Am. Chem. Soc.* **130**, 8502–8507 (2008).

[183] Weber, U. K. *et al.* Role of Interaction Anisotropy in the Formation and Stability of Molecular Templates. *Phys. Rev. Lett.* **100**, 156101 (2008).

[184] Velichko, Y. S., Stupp, S. I. & Olvera de la Cruz, M. Molecular simulation study of peptide amphiphile self-assembly. *J. Phys. Chem. B* **112**, 2326 (2008).

[185] Wilson, M. R. Molecular simulation of liquid crystals: Progress towards a better understanding of bulk structure and the prediction of material properties. *Chem. Soc. Rev.* **36**, 1881–1888 (2007).

[186] Doye, J. P. K. *et al.* Controlling crystallization and its absence: proteins, colloids and patchy models. *Phys. Chem. Chem. Phys.* **9**, 2197–2205 (2007).

[187] Jankowski, E. & Glotzer, S. C. A comparison of new methods for generating energy-minimizing configurations of patchy particles. *J. Chem. Phys.* **131**, 104104/1–8 (2009).

[188] Yeomans, J. *Statistical Mechanics of Phase Transitions* (Oxford Science Publications, Oxford, 1992).

[189] Hogyoku, M. Criticality found in a model for orientational ordering of protein arrays. *Adv. Biophys.* **34**, 55 (1997).

[190] Bianchi, E., Largo, J., Tartaglia, P., Zaccarelli, E. & Sciortino, F. Phase diagram of patchy colloids: Towards empty liquids. *Phys. Rev. Lett.* **97**, 168301 (2006).

[191] Zwanzig, R. First-order phase transition in a gas of long thin rods. *J. Chem. Phys.* **39**, 1714–1719 (1963).

[192] Krzakala, F., Tarzia, M. & Zdeborova, L. Lattice Model for Colloidal Gels and Glasses. *Phys. Rev. Lett.* **101**, 165702 (2008).

[193] Lu, P. J. *et al.* Gelation of particles with short-range attraction. *Nature* **453**, 499–U4 (2008).

[194] Zaccarelli, E. Colloidal gels: equilibrium and non-equilibrium routes. *J. Phys. Cond. Mat.* **19**, 323101 (2007).

[195] Lackinger, M., Griessl, S., Heckl, W. M., Hietschold, M. & Flynn, G. W. Self-assembly of trimesic acid at the liquid: Solid interface a study of solvent-induced polymorphism. *Langmuir* **21**, 4984–4988 (2005).

[196] Kelly, R. E. A. *et al.* Understanding the disorder of the DNA base cytosine on the Au(111) surface. *J. Chem. Phys.* **129**, 184707 (2008).

[197] Woodley, S. M. & Catlow, R. Crystal structure prediction from first principles. *Nature Materials* **7**, 937–946 (2008).

[198] Olson, A., Hu, Y. & Keinan, E. Chemical mimicry of viral capsid self-assembly. *Proc. Nat. Acad. Sci.* **104**, 20731 (2007).

[199] Davis, J. R. & Panagiotopoulos, A. Z. Monte Carlo simulations of amphiphilic nanoparticle self-assembly. *J. Chem. Phys.* **129**, 194706 (2008).

[200] Zhang, Z., Tang, Z., Kotov, N. A. & Glotzer, S. C. Simulations and analysis of self-assembly of CdTe nanoparticles into wires and sheets. *Nano. Lett.* **7**, 1670–1675 (2007).

[201] Adamczyk, Z., Barbasz, J. & Nattich, M. Particle assembly on patterned surfaces bearing circular (Dots) and rectangular (Stripes) surface features. *Langmuir* **24**, 1756–1762 (2008).

[202] Whitelam, S. & Geissler, P. L. Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles. *J. Chem. Phys.* **127**, 154101 (2007).

[203] Garcia-Sanchez, F., Valencia-Garcia, R., Martinez-Bejar, R. & Fernandez-Breis, J. T. An ontology, intelligent agent-based framework for the provision of semantic web services. *Expert Syst. Appl.* **36**, 3167–3187 (2009).

[204] Samanidou, E., Zschischang, E., Stauffer, D. & Lux, T. Agent-based models of financial markets. *Rep. Prog. Phys.* **70**, 409–450 (2007).

[205] Carley, K. M. Computational modeling for reasoning about the social behavior of humans. *Comput. Math. Organ. Theory* **15**, 47–59 (2009).

[206] Cioffi-Revilla, C. Simplicity and reality in computational modeling of politics. *Comput. Math. Organ. Theory* **15**, 26–46 (2009).

[207] Darbha, S., Rajagopal, K. R. & Tyagi, V. A review of mathematical models for the flow of traffic and some recent results. *Nonlinear Anal.-Theory Methods Appl.* **69**, 950–970 (2008).

[208] Furtado, L. & Copelli, M. Response of electrically coupled spiking neurons: A cellular automaton approach. *Phys. Rev. E* **73**, 011907 (2006).

[209] Schweitzer, F. Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals. *Nature* **441**, 815 (2006).

[210] Kier, L. A cellular automata model of bond interactions among molecules. *J. Chem. Inf. Comput. Sci.* **40**, 1285–1288 (2000).

[211] Kier, L. B., Seybold, P. G. & Cheng, C. K. *Modeling Chemical Systems Using Cellular Automata* (Spinger, New York, 2005).

[212] Seybold, P. G., O'Malley, M. J., Kier, L. B. & Cheng, C.-K. Cellular automata simulations of vapor-liquid equilibria. *Aust. J. Chem.* **59**, 865–868 (2006).

[213] Chen, T., Zhang, Z. & Glotzer, S. C. A precise packing sequence for self-assembled convex structures. *Proc. Nat. Acad. Sci.* **104**, 717–722 (2007).

[214] Nguyen, H. D., Reddy, V. S. & Brooks, C. L., III. Deciphering the kinetic mechanism of spontaneous self-assembly of icosahedral capsids. *Nano. Lett.* **7**, 338–344 (2007).

[215] Sear, R. P. Nucleation: theory and applications to protein solutions and colloidal suspensions. *J. Phys. Cond. Mat.* **19**, 033101 (2007).

[216] Wu, D., Chandler, D. & Smit, B. Electrostatic analogy for surfactant assemblies. *J. Phys. Chem.* **96**, 4077–4083 (1992).

[217] Tanner, M. A. & Wong, W. H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **82**, 528–540.

[218] Leardi, R. Genetic algorithms in chemistry. *J. Chromatogr. A* **1158**, 226–233 (2007).

[219] Gasteiger, J. & Zupan, J. Neural Networks in Chemistry. *Angew. Chem.* **32**, 503–527 (1993).

[220] Ponder, J. W. & Richards, F. M. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.* **8**, 1016 (1987).

[221] Wang, F. & Landau, D. P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**, 2050–2053 (2001).

[222] Kier, L. B., Cheng, C.-K. & Nelson, J. D. Models of solute aggregation using cellular automata. *Chem. Biodivers.* **6**, 396–401 (2009).

[223] Pothen, A., Simon, H. & Liou, K. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. A.* **11**, 430 (1990).

[224] Allinger, N. L. & co–workers. *MM3 Manual.* University of Georgia (1996).

[225] Canas-Ventura, M. E. *et al.* Coexistence of one- and two-dimensional supramolecular assemblies of terephthalic acid on Pd(111) due to self-limiting deprotonation. *J. Chem. Phys.* **125**, 184710 (2006).

[226] Martsinovich, N. & Troisi, A. Modeling the self-assembly of benzenedicarboxylic acids using Monte Carlo and molecular dynamics simulations. *J. Phys. Chem. C* **114**, 4376–4388 (2010).

[227] Lii, J. H. & Allinger, N. L. Directional hydrogen bonding in the MM3 force field: II. *J. Comput. Chem.* **19**, 1001–1016 (1998).

[228] Low, R. J. Measuring order and biaxiality. *Eur. J. Phys.* **23**, 111 (2002).

[229] Nicodemi, M., de Candia, A. & Coniglio, A. Aggregation of fibrils and plaques in amyloid molecular systems. *Phys. Rev. E* **80**, 041914 (2009).

[230] Fiedler, M. Algebraic connectivity of graphs. *Czech. Math. J.* **23**, 298 (1973).

[231] Anderson, E. *et al. LAPACK Users' Guide* (Society for Industrial and Applied Mathematics, Philadelphia, 1999), third edn.

[232] Phillips, J., Bruch, L. & Murphy, R. The two-dimensional Lennard-Jones system - Sublimation, Vaporization, and Melting. *J. Chem. Phys.* **75**, 5097–5109 (1981).

[233] de Miguel, E. & Vega, C. The global phase diagram of the Gay-Berne model. *J. Chem. Phys.* **117**, 6313–6322 (2002).

[234] Fernandez Toledano, J. C., Sciortino, F. & Zaccarelli, E. Colloidal systems with competing interactions: from an arrested repulsive cluster phase to a gel. *Soft Matter* **5**, 2390–2398 (2009).